The Effect of Mini-Batch Noise on the Implicit Bias of Adam

Matias D. Cattaneo*
Princeton University
cattaneo@princeton.edu

Boris Shigida*
Princeton University
bs1624@princeton.edu

Abstract

Adam and AdamW are standard optimizers in deep learning, but the choice of their momentum hyperparameters (β_1, β_2) is often not principled. We study how the interaction of these hyperparameters with batch size implicitly affects loss sharpness, relevant both to overfitting concerns in multi-epoch training and to post-training performance concerns after single-epoch pretraining. Our analysis predicts a batch-size-dependent reversal in the preferred choice of Adam's betas. In the low-noise, large-batch regime, increasing β_2 strengthens the implicit anti-regularization induced by memory, suggesting that choosing β_1 close to β_2 should improve generalization and model sensitivity. In contrast, in the high-noise, small-batch regime, mini-batch noise reverses this monotonicity: larger β_2 can compensate for the anti-regularizing memory term, making the default choice $\beta_1 \ll \beta_2$ theoretically well motivated. The predicted transition occurs at a batch-size scale controlled by the simple noise scale, closely related to the critical batch size. Experiments both in an overfitting regime and with single-epoch pretraining support these qualitative predictions.

1 Introduction

Adam [34] and its variants, including AdamW [43] and AdaFactor [61], are standard optimizers for modern deep learning tasks such as language-model training [9, 3, 67, 18]. Beyond the learning rate, training with these methods involves two consequential choices: the batch size b and Adam's momentum hyperparameters (β_1, β_2). These quantities jointly determine the optimizer's memory and the amount of mini-batch noise seen during training, yet their values are still largely set by convention.

The often appearing default traces back to Kingma and Ba [34], who recommend $(\beta_1, \beta_2) = (0.9, 0.999)$ based on a grid search. These values remain the defaults in widely used libraries such as PyTorch and Optax, and it is conventional wisdom that adaptive gradient methods work well with their default hyperparameters [62]. At the same time, recent large-model training practice often lowers β_2, for example to $\beta_2 = 0.95$, when using AdamW [9, 82, 79, 7, 67, 18], partly because this can improve training stability. These observations raise a basic question: should the preferred relationship between β_1, β_2, and batch size change across noise regimes?

The answer depends on the performance criterion. Large pretraining runs have often used only one pass, or less, over available data [8], where stability, optimization speed, and compute efficiency are primary concerns. In contrast, limited high-quality data can make multi-epoch training increasingly useful [69, 32], and parts of the post-training pipeline are explicitly multi-epoch and prone to overfitting [74]. In such regimes, including pretraining under data constraints [32], generalization quantities such as the train-validation gap are of interest.

Loss sharpness can be considered a mechanistic proxy for this type of generalization: there is a large body of work connecting flatter regions of the loss landscape to better generalization and showing the benefits of explicit or implicit sharpness penalization [29, 20, 38, 84, 33, 17, 42, 39, 76, 66, 41, 5, 63, 21, 59, 12]. (We provide some discussion of the evidence in Section 1.1.) Importantly, recent findings also suggest that sharpness during pretraining, even single-epoch, is closely related to downstream factors such as quantization degradation and catastrophic forgetting [72, 65]. This motivates asking how Adam's memory and mini-batch noise implicitly bias training toward or away from sharp regions of the loss landscape, irrespective of whether classical overfitting is a concern.

*Authors are listed alphabetically by last name.

We theoretically investigate how (β_1, β_2) and batch size b affect this implicit sharpness bias. Our analysis extends the framework of the implicit bias of memory [10] to isolate mini-batch noise effects in Adam. To our knowledge, this is the first theory explaining the batch-size-dependent reversal in Adam beta preferences through an implicit sharpness-bias mechanism.

Our main contributions are as follows.

1. In Section 2, we present a framework for finding and interpreting implicit bias terms during mini-batch training with an optimizer that has memory, meaning that the next iterate depends on the history of previous gradients. The exposition uses SGD with momentum as an illustrative example.

2. In Section 3, we apply this approach to mini-batch Adam. After removing memory and averaging over without-replacement mini-batch sampling, we derive an implicit-bias correction whose dominant terms can be interpreted through a sharpness proxy.

3. We show that mini-batch noise changes the monotonicity of the sharpness bias as a function of β_2 when β_1 is fixed. In the high-noise small-batch regime, larger β_2 can compensate for the anti-regularizing memory term, making the default ordering $\beta_1 \ll \beta_2$ suitable when overfitting or model sensitivity is a concern. In the low-noise large-batch regime, larger β_2 instead strengthens anti-regularization, suggesting that β_1 and β_2 should be chosen close to each other. As reviewed in Section 1.1, this prescription is consistent with a substantial body of empirical work.

4. We show that an analogous reversal appears when β_1 is swept with β_2 fixed at a common value such as 0.999. For large batches and full-batch training, larger β_1 is predicted to be better, consistent with Cattaneo et al. [12]; as mini-batch noise increases, this monotonicity reverses and again suggests $\beta_1 \ll \beta_2$ in small-batch regimes.

5. The transition scale in our simplified theory is controlled by the simple noise scale $\mathcal{B}_{\text{simple}}$, closely related to the critical batch size. In Section 4, small-scale language-model experiments in an overfitting regime along with a larger online pretraining run support these qualitative predictions.

1.1 Related Work

Tuning hyperparameters of Adam Ma et al. [45] investigate theoretically and empirically the qualitative features of full-batch Adam depending on (β_1, β_2), dividing possible training into three regimes (oscillations, spikes and divergence) and advocating for $\beta_1 = \beta_2$ for faster and smoother training. The latter prescription is consistent with our theory although we focus on different metrics (loss landscape sharpness / flatness), and we argue that increasing mini-batch noise changes the conclusions and recommendations. Relatedly, Zhao et al. [83] include Adam's (β_1, β_2) sweeps and find that if $\beta_1 = \beta_2$, Adam behaves similarly to signed momentum (Signum), and the recently common setting for language models $(\beta_1, \beta_2) = (0.9, 0.95)$ is close to this. For small batches, however, it is empirically observed to be beneficial to increase β_2 relative to β_1. In particular, Zhang et al. [81] recommend (based on empirical sweeps) taking smaller β_2 relative to β_1 if batch sizes are large and higher when batch sizes are small, exactly matching our theory-based prescription. There are other prior works advocating for that, and they use different principles [56, 47]. To the best of our knowledge, we provide the first theoretical argument based on generalization. Other works that have a substantial focus on (β_1, β_2) sweeps in Adam include Schmidt et al. [60], Orvieto and Gower [54], Wen et al. [73], Pagliardini et al. [55].

SDE approximations In the context of mini-batch noise, theoretical analysis of many optimizers often employs approximations by stochastic differential equations (SDEs). Relatedly, Jules et al. [30] use controlled thermal-like Langevin noise as a diagnostic probe of the low-loss landscape geometry of neural networks. In contrast, our work studies the implicit bias induced by the endogenous mini-batch noise and memory terms of Adam during training, and connects this effect to batch-size-dependent choices of (β_1, β_2). In particular, Zhou et al. [85], Xie et al. [77] approximate Adam and SGD with (different types of) SDEs, and use the escaping time from local minima to predict better generalization of SGD compared to Adam. The works Malladi et al. [46], Compagnoni et al. [13] also approximate Adam with SDEs under different assumptions and propose scaling rules for hyperparameters. Zhou et al. [86] focus on the advantages of decoupled weight decay for generalization. These works differ substantially from the present article in purpose, methods and assumptions; in particular, typically β_1 and β_2 are assumed to converge to 1 at certain rates as step size goes to zero, whereas we consider them fixed. In addition, we do

not assume that mini-batch noise in the gradients forms an i. i. d. random sequence (since we consider sampling without replacement), we are agnostic to its distribution, and we do not use distributional asymptotics.

Sharpness and generalization There has been a long history of relating flatter minima or loss regions to better generalization [24, 31, 29]. There has also been some criticism based on the sensitivity of standard sharpness measures to rescaling the network's parameters even if it does not change the network's outputs [16], which Kwon et al. [38] call the *scale dependency problem*. In response, different scale-invariant sharpness metrics have been introduced [78, 68, 58, 38]; however, empirical evidence is still mixed [1]. Numerous works have explored explicit sharpness penalization to improve generalization, of which we can only name a few [20, 38, 84, 33, 17, 42, 39, 76, 66, 41]. We study implicit, rather than explicit, penalization but otherwise our theory-based perspective is consistent with this literature. Although the non-adaptive sharpness metrics we find implicitly (anti-)penalized do have the scale dependency problem (along with most metrics in the related literature including Hessian-based ones), this does not invalidate any conclusions since penalizing or otherwise decreasing non-adaptive sharpness still often leads to better generalization.

Implicit bias A large strand of literature describes implicit biases of optimization algorithms by proving convergence to a max-margin solution [64, 52, 53, 57, 71, 22, 26, 27, 25, 23, 28, 51, 44, 70]. The implicit bias of weight decay in AdamW is tackled in Zhang et al. [80], Zhuang et al. [87], Andriushchenko et al. [2], Xie and Li [75], Kobayashi et al. [35] and others. Implicit regularization by biasing towards flatter minima at convergence is studied in Damian et al. [15], Arora et al. [4] besides works already listed. Most relatedly to our work, a large body of literature demonstrates implicit penalization of a gradient norm, using modified equations for SGD with or without momentum [5, 50, 63, 19, 36, 21, 59], for full-batch Adam [12], or using correction terms after removing memory [10]. Beneventano [6] studies the difference between SGD with or without replacement with similar methods. We build on and extend this literature.

2 Framework

This section introduces the framework used in the rest of the paper. We first specify the without-replacement mini-batch sampling model and the corresponding gradient-noise covariance. We then recall the memory-removal principle, which replaces an optimizer with memory by a memoryless iteration with an explicit correction term. Finally, we illustrate the method on SGD with momentum before applying it to Adam in Section 3.

Losses and gradients We assume there are $n+1$ batches in an epoch, each consisting of b samples, so that $N := (n+1)b$ samples are used in total. The kth *mini-batch loss* is defined by

$$\mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{b} \sum_{r=kb+1}^{kb+b} \ell_{\pi(r)}(\boldsymbol{\theta}), \quad k \in [0\,{:}\,n],$$

where $\{\ell_s\}_{s=1}^N$ are *per-sample losses* and $\pi\colon [1\,{:}\,N] \to [1\,{:}\,N]$ is a uniformly random permutation of the samples. The vector $\boldsymbol{\theta} \in \Theta$ collects all model parameters, and $\Theta \subset \mathbb{R}^{\dim \boldsymbol{\theta}}$ is the parameter domain of interest. The *full-batch loss* is the average of mini-batch losses:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n+1} \sum_{k=0}^{n} \mathcal{L}_k(\boldsymbol{\theta}) = \frac{1}{N} \sum_{r=1}^{N} \ell_r(\boldsymbol{\theta}).$$

We write the *loss gradient* as

$$\mathbb{R}^{\dim \boldsymbol{\theta}} \ni \boldsymbol{g} = (g_1, \dots, g_{\dim \boldsymbol{\theta}})^\mathsf{T} := \nabla \mathcal{L}(\boldsymbol{\theta}), \quad g_i := \partial_i \mathcal{L}(\boldsymbol{\theta}).$$

As usual, we omit the dependence on $\boldsymbol{\theta}$ when the point is fixed and clear from context.

Mini-batch noise, empirical covariance matrix We will denote for $k \in [0\,{:}\,n]$

$$d_k := (\mathcal{L}_k - \mathcal{L})(\boldsymbol{\theta})$$

the kth *mini-batch noise*. Since d_k is a function of $\boldsymbol{\theta}$, its derivatives $\partial_i d_k$ and $\partial_{ij} d_k$ are the corresponding gradient and Hessian noise. Further, we define the *empirical covariance matrix* $\boldsymbol{\Sigma}$ of per-sample gradients:

$$\Sigma_{ij} := \frac{1}{(n+1)b} \sum_{p=1}^{(n+1)b} \partial_i(\ell_p - \mathcal{L})(\boldsymbol{\theta})\partial_j(\ell_p - \mathcal{L})(\boldsymbol{\theta}), \quad \boldsymbol{\Sigma} = (\Sigma_{ij})_{i,j=1}^{\dim \boldsymbol{\theta}} \in \mathbb{R}^{\dim \boldsymbol{\theta} \times \dim \boldsymbol{\theta}}.$$

Memory Removal An optimization algorithm has *memory* if its update depends on the history of previous iterates, not only on the current iterate. Consider a general iteration of this form:

$$\boldsymbol{\theta}_{t+1} = \underbrace{\boldsymbol{\theta}_t - \eta \boldsymbol{F}_t(\boldsymbol{\theta}_t, \dots, \boldsymbol{\theta}_0)}_{\text{depends on the whole history } \boldsymbol{\theta}_t, \dots, \boldsymbol{\theta}_0}. \tag{1}$$

The memory-removal result of Cattaneo and Shigida [10] converts it into a memoryless iteration

$$\tilde{\boldsymbol{\theta}}_{t+1} = \underbrace{\tilde{\boldsymbol{\theta}}_t - \eta\, \mathbf{Main}_t(\tilde{\boldsymbol{\theta}}_t) - \eta^2\, \mathbf{Corr}_t(\tilde{\boldsymbol{\theta}}_t)}_{\text{only depends on } \tilde{\boldsymbol{\theta}}_t \text{ (no memory)}}, \tag{2}$$

where $\mathbf{Main}_t$ is the update obtained by freezing the whole history at one point and $\mathbf{Corr}_t$ is the correction that compensates for this replacement. The original and memoryless trajectories stay globally $O(\eta^2)$-close for $O(\eta^{-1})$ iterations: for any "physical time" horizon $T > 0$, there is a constant C such that

$$\max_{t \in [0:\lfloor T/\eta \rfloor]} \left\| \boldsymbol{\theta}_t - \tilde{\boldsymbol{\theta}}_t \right\|_\infty \le C\eta^2, \tag{3}$$

provided that the *main term* $\mathbf{Main}_t(\boldsymbol{\theta}) \in \mathbb{R}^{\dim \boldsymbol{\theta}}$ and the *correction term* $\mathbf{Corr}_t(\boldsymbol{\theta}) \in \mathbb{R}^{\dim \boldsymbol{\theta}}$ are chosen as

$$\mathbf{Main}_t(\boldsymbol{\theta}) := \boldsymbol{F}_t(\boldsymbol{\theta}, \dots, \boldsymbol{\theta}), \quad \mathrm{Corr}_{t,r}(\boldsymbol{\theta}) := \sum_{k=1}^{t} \frac{\partial F_{t,r}}{\partial \boldsymbol{\theta}_{t-k}}(\boldsymbol{\theta})^\intercal \sum_{s=t-k}^{t-1} \boldsymbol{F}_s(\boldsymbol{\theta}). \tag{4}$$

Removing memory is the first step in our analysis. We then interpret the correction terms in the resulting memoryless iteration, asking whether they penalize or anti-penalize sharpness. We now illustrate the procedure on a simple algorithm with memory.

2.1 Warm-up: SGD with Momentum

Consider mini-batch SGD with momentum, written in the form Eq. (1) with $F_t(\boldsymbol{\theta}_t, \dots, \boldsymbol{\theta}_0) := \sum_{k=0}^{t} \beta^{t-k} \nabla \mathcal{L}_k(\boldsymbol{\theta}_k)$. This example introduces the main ideas behind the framework; see Cattaneo and Shigida [11] for a fine-grained analysis of this specific algorithm.

The memory removal technique just described (formally Theorem C.1) gives an approximation (2) with

$$\begin{aligned} \mathbf{Main}_t(\boldsymbol{\theta}) &= \sum_{k=0}^{t} \beta^{t-k} \nabla \mathcal{L}_k(\boldsymbol{\theta}), \\ \mathbf{Corr}_t(\boldsymbol{\theta}) &= \beta \sum_{q=0}^{t-1} \beta^q \sum_{l=1}^{q+1} \sum_{q_1=0}^{t-l} \beta^{q_1} \nabla^2 \mathcal{L}_{t-1-q}(\boldsymbol{\theta}) \nabla \mathcal{L}_{t-l-q_1}(\boldsymbol{\theta}). \end{aligned} \tag{5}$$

The approximating algorithm does not have memory, so $\mathbf{Main}_t$ and $\mathbf{Corr}_t$ only depend on one point, which is already a significant simplification. However, in this form these expressions are still very complex and their analysis appears impossible. The next step (due to Smith et al. [63]), is to put $t = n$ and take the average $\mathbb{E}_\pi$ over all permutations of samples. This gives the average one-epoch correction at the current point $\boldsymbol{\theta}$, removing the accidental dependence on a particular batch order. The average of the main term is easy to compute:

$$\mathbb{E}_\pi \mathbf{Main}_n(\boldsymbol{\theta}) = \sum_{k=0}^{n} \beta^{n-k} \mathbb{E}_\pi \nabla \mathcal{L}_k(\boldsymbol{\theta}) = \sum_{k=0}^{n} \beta^{n-k} \boldsymbol{g} = \frac{1-\beta^{n+1}}{1-\beta} \boldsymbol{g} = \frac{1+o_n(1)}{1-\beta} \boldsymbol{g},$$

where $o_n(1)$ denotes terms that decay to zero as $n \to \infty$ (exponentially fast). After some similar algebra, we also find the average correction:

$$\mathbb{E}_\pi \mathbf{Corr}_n(\boldsymbol{\theta}) = \frac{\beta + o_n(1)}{2(1-\beta)^3} \nabla \|\boldsymbol{g}\|^2 + \frac{\beta + o_n(1)}{2(1-\beta)^2(1+\beta)} \nabla \left(\frac{\operatorname{tr} \boldsymbol{\Sigma}}{b} \right).$$

In other words,

$$\mathbb{E}_\pi \mathbf{Main}_n(\boldsymbol{\theta}) + \eta \, \mathbb{E}_\pi \mathbf{Corr}_n(\boldsymbol{\theta}) = \frac{1}{1-\beta} \nabla \left((1 + o_n(1)) \mathcal{L} + \eta \frac{\beta + o_n(1)}{2(1-\beta)^2} \|\boldsymbol{g}\|^2 + \eta \frac{\beta + o_n(1)}{2(1-\beta)(1+\beta)} \frac{\operatorname{tr} \boldsymbol{\Sigma}}{b} \right).$$

This expression is non-random and is much easier to analyze. In the right-hand side, we see a modified loss with two correction terms:

- implicit regularization by memory $\eta \frac{\beta + o_n(1)}{2(1-\beta)^2} \|\boldsymbol{g}\|^2$ (present already in the full-batch case), and
- implicit regularization by stochasticity $\eta \frac{\beta + o_n(1)}{2(1-\beta)(1+\beta)} \frac{\operatorname{tr} \boldsymbol{\Sigma}}{b}$ (appearing as a result of mini-batch noise).

The first term implicitly penalizes the squared norm of the gradient, which is a first-order approximation of (non-adaptive) ℓ_2 sharpness [20]: for small ρ,

$$\max_{\|\boldsymbol{\epsilon}\| \le \rho} \mathcal{L}(\boldsymbol{\theta} + \boldsymbol{\epsilon}) - \mathcal{L}(\boldsymbol{\theta}) \approx \max_{\|\boldsymbol{\epsilon}\| \le \rho} \boldsymbol{g}^\mathsf{T} \boldsymbol{\epsilon} = \rho \|\boldsymbol{g}\|.$$

The second term implicitly penalizes gradient noise variance

$$\operatorname{tr} \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \frac{1}{(n+1)b} \sum_{p=1}^{(n+1)b} \left\| \nabla(\ell_p - \mathcal{L})(\boldsymbol{\theta}) \right\|^2 = \frac{nb + b - 1}{n} \sum_i \mathbb{E}_\pi (\partial_i d_0)^2,$$

also related to flatness of the loss landscape and observed to be predictive of generalization [29].

Therefore, penalizing both terms is predictive of moving toward flatter regions of the loss landscape and often better generalization. Following tradition, we classify them as "implicit regularization".

2.2 Summary

Our approach interprets implicit biases of mini-batch versions of optimization algorithms with memory using the following three steps.

1. **Removing memory:** use the memory-removal technique to approximate the algorithm with memory by a memoryless iteration.
2. **Calculating the average correction terms:** take expectation $\mathbb{E}_\pi \mathbf{Corr}_n(\boldsymbol{\theta})$ to remove dependence on a particular mini-batch order, and potentially make other simplifications without qualitatively changing the situation.
3. **Interpretation:** interpret the terms in the resulting expression, especially connecting to known sharpness/flatness or generalization measures.

This framework gives a way to study how mini-batch noise influences, on average, the implicit bias of memory in complex optimization algorithms used for deep learning.

3 Mini-Batch Noise in Adam

We now apply the framework from Section 2 to Adam. The section has three steps. First, we remove memory and obtain a memoryless approximation with an explicit correction term. Second, we average this correction over without-replacement mini-batch order and decompose it into a full-batch term and five mini-batch-noise terms. Third, we interpret the dominant terms through a sharpness proxy and derive directional predictions for how the preferred betas change with batch size.

We use Adam in its equivalent one-variable form, obtained by eliminating the first- and second-moment variables.

Definition 3.1 (Adam [34])**.** For numerical hyperparameters $\epsilon > 0$ and $\beta_1, \beta_2 \in (0,1)$, Adam can be written for each coordinate $j \in [1 : \dim \boldsymbol{\theta}]$ as

$$\theta_{t+1,j} = \theta_{t,j} - \eta \frac{\sum_{k=0}^{t} \mu_{t,k} \partial_j \mathcal{L}_k(\boldsymbol{\theta}_k)}{\sqrt{\sum_{k=0}^{t} \nu_{t,k} |\partial_j \mathcal{L}_k(\boldsymbol{\theta}_k)|^2 + \epsilon}},$$
$$\text{where} \quad \mu_{t,k} := \frac{\beta_1^{t-k}(1-\beta_1)}{1-\beta_1^{t+1}}, \quad \nu_{t,k} := \frac{\beta_2^{t-k}(1-\beta_2)}{1-\beta_2^{t+1}}, \quad k \in [0:t], t \in \mathbb{Z}_{\geq 0},$$

with arbitrary initial parameter $\boldsymbol{\theta}_0 \in \mathbb{R}^{\dim \boldsymbol{\theta}}$.

3.1 Step 1: Removing Memory

An application of the memory removal technique to the case of mini-batch Adam provides the following result whose full version is Theorem B.1.

Theorem 3.2 (Memory removal, simplified version)**.** *The iteration $\{\boldsymbol{\theta}_t\}_{t=0}^{\infty}$ given by Adam (Definition 3.1) is $O(\eta^2)$-close for $O(\eta^{-1})$ steps in the sense of bound (3) to the iteration $\{\tilde{\boldsymbol{\theta}}_t\}_{t=0}^{\infty}$ given by*

$$\tilde{\boldsymbol{\theta}}_{t+1} = \tilde{\boldsymbol{\theta}}_t - \eta \, \boldsymbol{Main}_t(\tilde{\boldsymbol{\theta}}_t) - \eta^2 \, \boldsymbol{Corr}_t(\tilde{\boldsymbol{\theta}}_t), \quad \tilde{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0,$$

where

$$Main_{t,j}(\boldsymbol{\theta}) := \frac{\sum_{k=0}^{t} \mu_{t,k} \partial_j \mathcal{L}_k(\boldsymbol{\theta})}{\sqrt{\sum_{k=0}^{t} \nu_{t,k} |\partial_j \mathcal{L}_k(\boldsymbol{\theta})|^2 + \epsilon}},$$

and the full expression for $\boldsymbol{Corr}_t(\boldsymbol{\theta})$ is deferred to (11) in Section B due to its length.

The proof follows from the general result in Cattaneo and Shigida [10], and is given in Section C. The terms $\mathbf{Main}_t(\boldsymbol{\theta})$ and $\mathbf{Corr}_t(\boldsymbol{\theta})$ are complex and difficult to interpret. Thus, we proceed to the next step to simplify the analysis.

3.2 Step 2: Calculating the Average Correction Terms

In this step, we put $t = n$, expand $\mathrm{Corr}_{n,j}(\boldsymbol{\theta})$ up to degree-2 monomials in noise derivatives, and calculate the average of the result with respect to permutations of samples. The expansion is local and second order in mini-batch noise. It is intended to capture directional effects in regimes where these terms dominate the omitted higher-order corrections, rather than to give a uniformly accurate quantitative approximation for every extremely small batch size. Recall that d_k, $\partial_i d_k$, $\partial_{ij} d_k$ denote the mini-batch noise and its partial derivatives. Accordingly, we will use the notation $O(d^p)$ to mean "terms of order at least p in (derivatives of) noise". For example, all terms of the form $(\partial_{ij} d_k)(\partial_i d_k)$ are $O(d^2)$ and all terms of the form $(\partial_{ijl} d_k)(\partial_{ij} d_k)(\partial_l d_k)$ are $O(d^3)$.

The following is a simplified combination of Theorems B.2 and B.3, proven in Sections D and E. The appendix keeps the nonzero-ϵ expressions; the main text presents the cleaner small-ϵ form.

Theorem 3.3 (Mini-batch noise expansion of the memoryless dynamics, simplified version)**.** *The expectation $\mathbb{E}_\pi$ of the correction term with respect to the uniform law on all permutations $[1 : (n+1)b] \to [1 : (n+1)b]$ satisfies, up to small ϵ corrections and finite-epoch terms of order $o_n(b^{-1})$,*

$$\begin{aligned} |g_j| \mathbb{E}_\pi Corr_{n,j} = {} & FB_j(\beta_1, \beta_2) + MBN_{1,j}(\beta_1, \beta_2) + MBN_{2,j}(\beta_1, \beta_2) \\ & + MBN_{3,j}(\beta_1, \beta_2) + MBN_{4,j}(\beta_1, \beta_2) + MBN_{5,j}(\beta_1, \beta_2) + O(d^3), \end{aligned} \tag{6}$$

where the full-batch correction is given by

$$FB_j(\beta_1, \beta_2) := \left(\frac{\beta_1}{1-\beta_1} - \frac{\beta_2}{1-\beta_2} \right) \partial_j \|\boldsymbol{g}\|_1, \tag{7}$$

Table 1: Interpretation of the terms in the mini-batch-noise expansion.

Term	Role	Treatment in the main interpretation
FB_j	Full-batch memory correction; for $\beta_1 < \beta_2$, it anti-penalizes the non-adaptive ℓ_∞ sharpness proxy.	Kept; this is the baseline anti-regularizing term that mini-batch noise competes with.
$\mathrm{MBN}_{1,j}$	Diagonal noise-to-signal correction proportional to Σ_{jj}/g_j^2.	Kept; it contributes to the sharpness-bias coefficient after the simple-noise-scale approximation.
$\mathrm{MBN}_{2,j}$	Cross-coordinate correction involving $\sum_i \partial_j \lvert g_i \rvert \Sigma_{ii}/g_i^2$.	Kept; after replacing per-coordinate noise-to-signal ratios by their global average, it has the same sharpness-proxy structure as $\mathrm{MBN}_{1,j}$.
$\mathrm{MBN}_{3,j}$	Covariance-derivative term involving $\partial_i \Sigma_{jj}$.	Treated as sign-neutral for the sharpness direction; Section F.2 gives the heuristic argument.
$\mathrm{MBN}_{4,j}$, $\mathrm{MBN}_{5,j}$	Remaining covariance-derivative and off-diagonal covariance terms.	Neglected in the main interpretation because their coefficients are small in the beta ranges studied; see Lemma F.1.

and five mini-batch noise corrections are given by

$$
\begin{aligned}
\mathit{MBN}_{1,j}(\beta_1,\beta_2) &:= \frac{1}{b} C_1(\beta_1,\beta_2)\partial_j \|\boldsymbol{g}\|_1 \frac{\Sigma_{jj}}{g_j^2},\\
\mathit{MBN}_{2,j}(\beta_1,\beta_2) &:= \frac{1}{b} C_2(\beta_1,\beta_2)\sum_i \partial_j |g_i| \frac{\Sigma_{ii}}{g_i^2},\\
\mathit{MBN}_{3,j}(\beta_1,\beta_2) &:= \frac{1}{b} C_3(\beta_1,\beta_2)\frac{\operatorname{sign} g_j}{|g_j|}\sum_i \operatorname{sign} g_i\, \partial_i \Sigma_{jj},\\
\mathit{MBN}_{4,j}(\beta_1,\beta_2) &:= \frac{1}{b} C_4(\beta_1,\beta_2)\sum_i \frac{1}{|g_i|}\partial_j \Sigma_{ii},\\
\mathit{MBN}_{5,j}(\beta_1,\beta_2) &:= \frac{1}{b} C_5(\beta_1,\beta_2)\frac{\operatorname{sign} g_j}{|g_j|}\sum_i \frac{\partial_i g_j}{|g_i|}\Sigma_{ij},
\end{aligned}
\tag{8}
$$

with Σ_{ij} denoting the (i,j)th component of the empirical per-sample gradient covariance matrix $\boldsymbol{\Sigma}$, and the values of $\{C_k(\beta_1,\beta_2)\}_{k=1}^5$ deferred to Eq. (13) in Section B due to their length.

For very small batches, these degree-2 monomials may not be enough for an accurate quantitative approximation. Our predictions below are therefore directional: they describe the sign and monotonicity of the sharpness-bias coefficient, not the exact optimal performance boundaries.

3.3 Step 3: Interpretation

We need to analyze each term in the right-hand side of Eq. (6). Since the theorem is written after multiplying the correction by $|g_j|$, all parts of the correction term have the same preconditioning $|g_j|^{-1}$ as the full-batch Adam does. Because the sign of the effect depends on the region of hyperparameter space, we focus on two one-dimensional sweeps. In each case, one beta is fixed at a common value and we ask which value of the other beta is preferred under the sharpness proxy discussed in the introduction. Specifically, we study how to set β_2 when β_1 is fixed at 0.9, and how to set β_1 when β_2 is fixed at the default value 0.999.

How to set β_2 if β_1 is fixed We start with the setting where β_1 is fixed at its default value 0.9 and β_2 varies in the interval $[0.9, 1)$:

$$\beta_1 = 0.9, \quad \text{seeking preferred } \beta_2 \in [0.9, 1).$$

The terms $\mathrm{MBN}_{4,j}(\beta_1,\beta_2)$ and $\mathrm{MBN}_{5,j}(\beta_1,\beta_2)$ are easiest to handle: Lemma F.1 shows that their coefficients are small compared to the dominant terms in this beta range. In addition, we argue in Section F.2 that $\mathrm{MBN}_{3,j}(\beta_1,\beta_2)$ is sign-neutral for our sharpness-direction interpretation.

We are left with the sum of three terms: $\mathrm{FB}_j(\beta_1,\beta_2)$, $\mathrm{MBN}_{1,j}(\beta_1,\beta_2)$ and $\mathrm{MBN}_{2,j}(\beta_1,\beta_2)$. The full-batch term $\mathrm{FB}_j(\beta_1,\beta_2)$ anti-penalizes, when $\beta_1<\beta_2$, the 1-norm of the gradient. This is a first-order approximation of non-adaptive ℓ_∞-sharpness: for small ρ,

$$\max_{\|\boldsymbol{\epsilon}\|_\infty\le\rho}\mathcal{L}(\boldsymbol{\theta}+\boldsymbol{\epsilon})-\mathcal{L}(\boldsymbol{\theta})\approx\max_{\|\boldsymbol{\epsilon}\|_\infty\le\rho}\boldsymbol{g}^\mathsf{T}\boldsymbol{\epsilon}=\rho\|\boldsymbol{g}\|_1.$$

Thus, we can refer to this term as anti-regularization, same as in the setting with zero noise [12]. The term containing $C_1(\beta_1,\beta_2)$ (it can be checked that it is positive in our setting) provides regularization: it also penalizes the ℓ_1 gradient norm although the magnitude of this penalization in each component j depends on the per-component noise-to-signal ratio Σ_{jj}/g_j^2.

The term $\mathrm{MBN}_{2,j}(\beta_1,\beta_2)$ is more complicated, so we make an explicit simple-noise-scale approximation. At the current point $\boldsymbol{\theta}$, we replace the per-coordinate noise-to-signal ratios Σ_{ii}/g_i^2 by a single global average, the "simple noise scale" $\mathcal{B}_{\mathrm{simple}}$ from [48]:

$$\mathcal{B}_{\mathrm{simple}}:=\frac{\operatorname{tr}\boldsymbol{\Sigma}}{\sum_j g_j^2}=\frac{\operatorname{tr}\boldsymbol{\Sigma}}{\|\boldsymbol{g}\|^2}.$$

This approximation discards per-coordinate variation in Σ_{ii}/g_i^2, but it preserves the global scale of mini-batch noise relative to the full-batch gradient. After this replacement, the terms $\mathrm{MBN}_{1,j}(\beta_1,\beta_2)$ and $\mathrm{MBN}_{2,j}(\beta_1,\beta_2)$ have the same sharpness-proxy structure up to their coefficients: $C_1(\beta_1,\beta_2)\partial_j\|\boldsymbol{g}\|_1 b^{-1}\mathcal{B}_{\mathrm{simple}}$ and $C_2(\beta_1,\beta_2)\partial_j\|\boldsymbol{g}\|_1 b^{-1}\mathcal{B}_{\mathrm{simple}}$ respectively.

Under this approximation, the dominant terms in (6) provide implicit (anti-)penalization of an approximate non-adaptive sharpness measure with coefficient $C_{\mathrm{total}}\big(\beta_1,\beta_2,b^{-1}\mathcal{B}_{\mathrm{simple}}\big)$, where the function $(0,+\infty)\ni\lambda\mapsto C_{\mathrm{total}}\big(\beta_1,\beta_2,\lambda\big)$ is defined by

$$C_{\mathrm{total}}(\beta_1,\beta_2,\lambda):=\frac{\beta_1}{1-\beta_1}-\frac{\beta_2}{1-\beta_2}+\{C_1(\beta_1,\beta_2)+C_2(\beta_1,\beta_2)\}\lambda. \tag{9}$$

If $C_{\mathrm{total}}(\beta_1,\beta_2,b^{-1}\mathcal{B}_{\mathrm{simple}})>0$, this can be interpreted as regularization, otherwise as anti-regularization.

It remains to analyze how this coefficient depends on $\lambda>0$. We use the following fact whose full version is Proposition G.1.

Proposition 3.4 (Monotonicity of $C_{\mathrm{total}}(0.9,\beta_2,\lambda)$, simplified version)**.** *If $\lambda\ge 0.5082$, the function $C_{total}(0.9,\beta_2,\lambda)$ is increasing in $\beta_2\in[0.9,1)$. If $0<\lambda<0.494$, it is decreasing in $\beta_2\in[0.9,1)$.*

We obtain the following prediction. For fixed $\beta_1=0.9$ and $\beta_2\in[0.9,1)$, the quantity $b^{-1}\mathcal{B}_{\mathrm{simple}}$ controls the monotonicity of the approximate sharpness-bias coefficient. If this quantity is significantly below 0.5, equivalently if $b\gg 2\mathcal{B}_{\mathrm{simple}}$, the coefficient decreases with β_2: higher β_2 means weaker sharpness penalization, often predicting worse generalization. If the quantity is significantly above 0.5, equivalently if $b\ll 2\mathcal{B}_{\mathrm{simple}}$, the coefficient increases with β_2: higher β_2 means stronger sharpness penalization, often predicting better generalization:

$$\begin{aligned} b\gg 2\mathcal{B}_{\mathrm{simple}} \quad&\Rightarrow\quad C_{\mathrm{total}}\big(0.9,\beta_2,b^{-1}\mathcal{B}_{\mathrm{simple}}\big)\searrow \text{ in } \beta_2,\\ b\ll 2\mathcal{B}_{\mathrm{simple}} \quad&\Rightarrow\quad C_{\mathrm{total}}\big(0.9,\beta_2,b^{-1}\mathcal{B}_{\mathrm{simple}}\big)\nearrow \text{ in } \beta_2.\end{aligned}$$

Theoretically, the transition happens very quickly around the point where the batch size is $2\mathcal{B}_{\mathrm{simple}}$, although simplifications that we used make the theoretical transition quicker than it is in practice.

How to set β_1 if β_2 is fixed Next, we consider the complementary one-dimensional sweep, where β_2 is fixed at a default value and β_1 varies. This sweep is less common in practice, but it is still useful for understanding the full picture:

$$\beta_2=0.999,\quad \text{seeking preferred } \beta_1\in[0.9,1).$$

The following simplified variant of Proposition G.2 describes this situation.

Proposition 3.5 (Monotonicity of $C_{\text{total}}(\beta_1, 0.999, \lambda)$, simplified version)**.** *If $\lambda \geq 1.002$, the function $C_{total}(\beta_1, 0.999, \lambda)$ is strictly decreasing in $\beta_1 \in [0.9, 1)$. If $0 < \lambda < 0.995$, it is strictly increasing in $\beta_1 \in [0.9, 1)$.*

In this case, if $b^{-1}\mathcal{B}_{\text{simple}}$ is much larger than one ($b \ll \mathcal{B}_{\text{simple}}$), the approximate sharpness-bias coefficient decreases with β_1. Thus, larger β_1 weakens the bias toward flatter regions, often predicting worse generalization, and the sharpness proxy favors taking β_1 as low as stability allows (for example 0.9). If $b^{-1}\mathcal{B}_{\text{simple}}$ is much smaller than one ($b \gg \mathcal{B}_{\text{simple}}$), the coefficient increases with β_1. In that low-noise regime, the sharpness proxy favors moving β_1 upward, with $\beta_1 = \beta_2$ as a natural first choice when training remains stable.

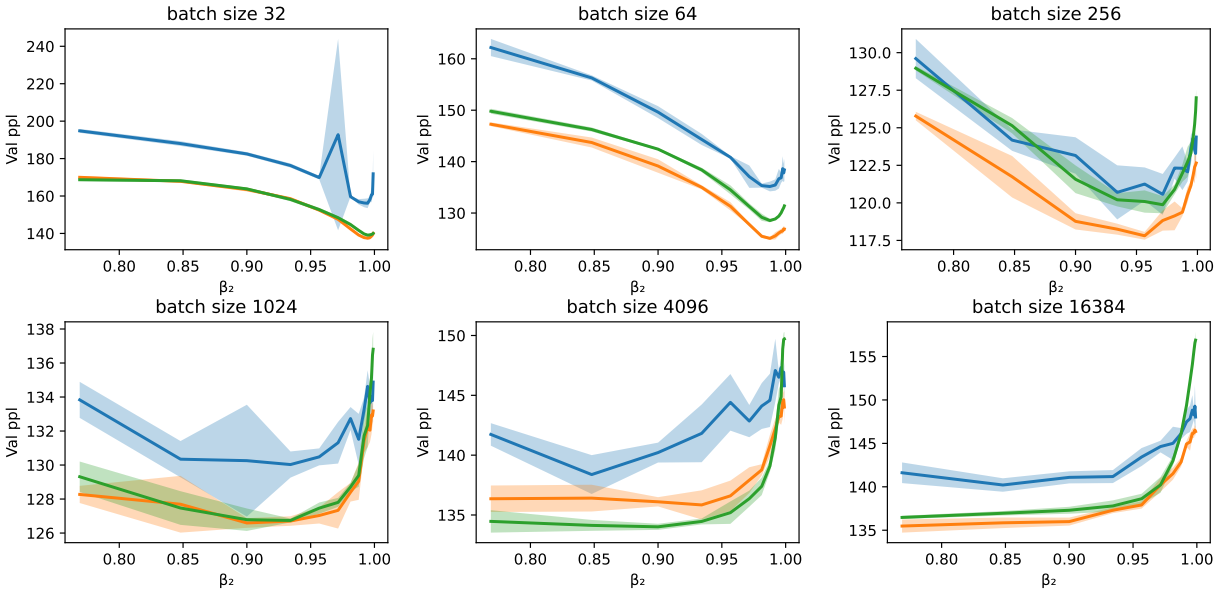


Figure 1: Minimal validation perplexity (before overfitting) of a small Transformer trained with Adam on WikiText-2 with different batch sizes, learning rates $\{10^{-3}, 10^{-3.5}, 10^{-4}\}$, $\beta_1 = 0.9$ (averaged over three iterations).

Takeaway If the batch size is much smaller than $\mathcal{B}_{\text{simple}}$, take β_1 much smaller than β_2 (e. g., the default values $\beta_1 = 0.9$, $\beta_2 = 0.999$ are a reasonable first choice). If the batch size is much larger than $2\mathcal{B}_{\text{simple}}$, take $\beta_1 = \beta_2$ (e. g., $\beta_1 = \beta_2 = 0.9$ is a natural first choice).

Because the derivation uses simplifications, the predictions are directional rather than precisely quantitative. In particular, the main conclusion is about the scale of the transition, controlled by $\mathcal{B}_{\text{simple}}$ (or $2\mathcal{B}_{\text{simple}}$), which is not difficult to estimate [48]. Thus, the rule of thumb can guide practical choices of β_1 and β_2 without requiring very large grids.

4 Experiments

Small Transformer overfitting on a small dataset We train a small Transformer from [14] on WikiText-2 [49] following Kunstner et al. [37]. We fix the default value $\beta_1 = 0.9$, and sweep β_2 for different batch sizes and learning rates. Running sufficiently many epochs to let the model overfit, we plot the minimal validation perplexity achieved depending on β_2. The results in Fig. 1 show that in small-batch Adam, larger β_2 mostly helps the model generalize better (decreases minimal validation perplexity), and this behavior smoothly transitions into the opposite as the batch size increases. To track the mechanism, we also plot in Fig. 2 non-adaptive ℓ_∞ sharpness evaluated exactly, along with its approximation $\|\boldsymbol{g}\|_1$ for a selected learning rate and two (small and large) batch sizes. Additional figures, including a sweep of β_1 at a fixed $\beta_2 = 0.999$, are provided in Section A.

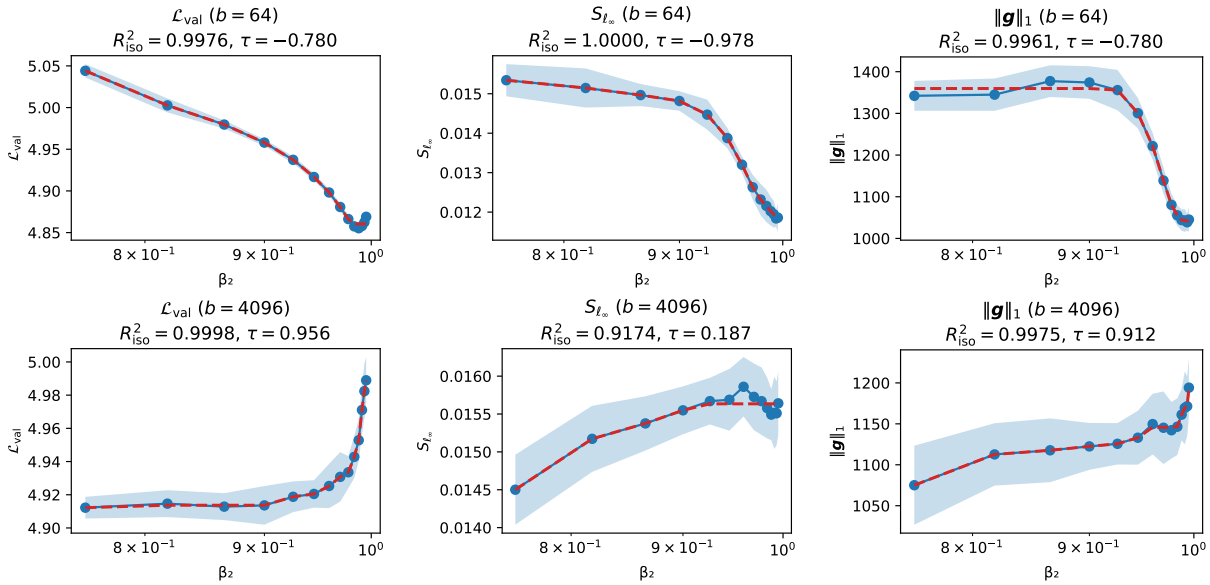


Figure 2: Validation loss, ℓ_∞-sharpness, ℓ_1-norm of the gradient as a function of β_2 at the median epoch of overfitting, for a small Transformer trained with Adam on WikiText-2 at a small (top) and large (bottom) batch size (averaged across at least 16 iterations). Red dashed line denotes an isotonic regression fit; τ denotes Kendall's tau.

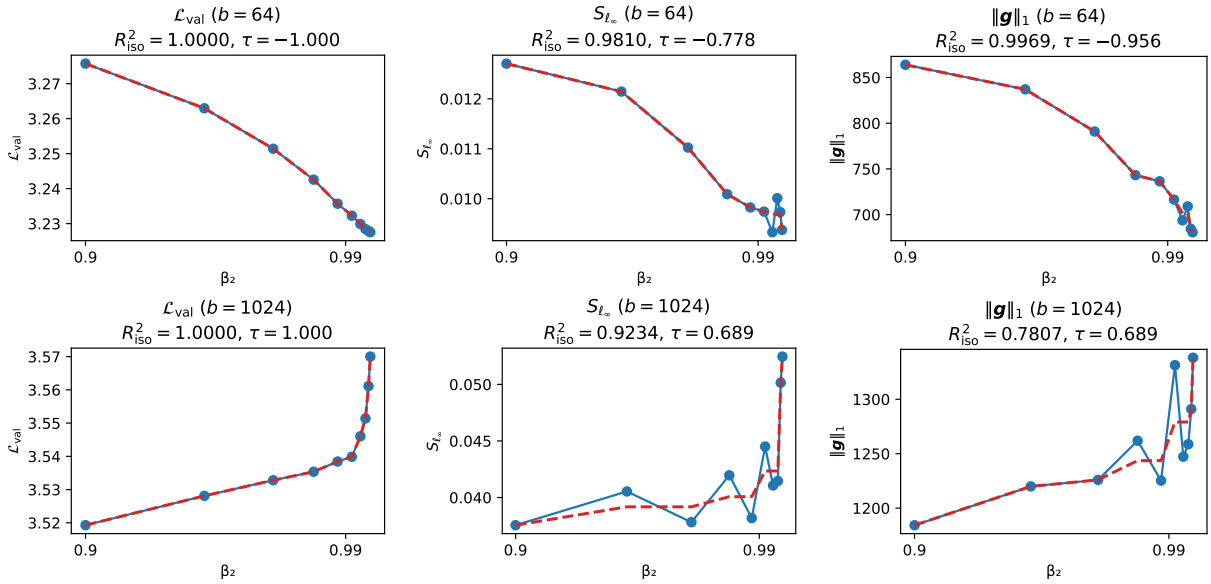


Figure 3: Validation loss, ℓ_∞-sharpness, ℓ_1-norm of the gradient as a function of β_2 at the end of training, for a 280 M-parameter Llama3 trained with AdamW on DCLM at a small (top) and large (bottom) batch size matched on total token budget. Red dashed line denotes an isotonic regression fit; τ denotes Kendall's tau.

Llama-3 on DCLM We train a modern Llama3 [18] model with 280 M parameters (no weight tying) on DCLM [40] with sequence length 4096 for about 25 B tokens (90 tokens per parameter). The learning rate is warmed up linearly for the first 10% of the training run and then decayed at a cosine schedule to 10% of the peak value. The optimizer is AdamW [43] with decoupled weight decay $\lambda = 0.1$ applied to all but the normalization parameters. No batches of training data are repeated, and the number of iterations is chosen to match the token count. In Fig. 3, we plot the loss, ℓ_∞ sharpness and 1-norm of the full-batch gradient

on a fixed set of 1024 validation sequences (full-batch calculation on the training data is infeasible), for a fixed peak learning rate and two (small and large) pretraining batch sizes, sweeping β_2. We observe the familar monotonicity reversion. Although train/validation performance gap is not applicable, even in single-epoch training the trends may be important for post-training or post-quantization performance [72, 65].

Acknowledgments

Cattaneo gratefully acknowledges financial support from the National Science Foundation through DMS-2210561 and SES-2241575. We acknowledge the Princeton Research Computing resources, coordinated by the Princeton Institute for Computational Science and Engineering (PICSciE) and the Office of Information Technology's Research Computing.

References

[1] M. Andriushchenko, F. Croce, M. Müller, M. Hein, and N. Flammarion. A modern look at the relationship between sharpness and generalization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023. URL https://proceedings.mlr.press/v202/andriushchenko23a.html.

[2] M. Andriushchenko, F. D'Angelo, A. Varre, and N. Flammarion. Why do we need weight decay in modern deep learning?, 2024. URL https://openreview.net/forum?id=RKh7DI23tz.

[3] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[4] S. Arora, Z. Li, and A. Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022. URL https://proceedings.mlr.press/v162/arora22a.html.

[5] D. Barrett and B. Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3q5IqUrkcF.

[6] P. Beneventano. On the trajectories of sgd without replacement. *arXiv preprint arXiv:2312.16143*, 2023.

[7] S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[10] M. D. Cattaneo and B. Shigida. How memory in optimization algorithms implicitly modifies the loss. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2qd4lpXz7u.

[11] M. D. Cattaneo and B. Shigida. Modified loss of momentum gradient descent: Fine-grained analysis. *arXiv preprint arXiv:2509.08483*, 2025.

[12] M. D. Cattaneo, J. M. Klusowski, and B. Shigida. On the implicit bias of Adam. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v235/cattaneo24a.html.

[13] E. M. Compagnoni, T. Liu, R. Islamov, F. N. Proske, A. Orvieto, and A. Lucchi. Adaptive methods through the lens of SDEs: Theoretical insights on the role of noise. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ww3CLRhF1v.

[14] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://api.semanticscholar.org/CorpusID:57759363.

[15] A. Damian, T. Ma, and J. D. Lee. Label noise sgd provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e6af401c28c1790eaef7d55c92ab6ab6-Paper.pdf.

[16] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. PMLR, 2017. URL https://proceedings.mlr.press/v70/dinh17b.html.

[17] J. Du, D. Zhou, J. Feng, V. Tan, and J. T. Zhou. Sharpness-aware training for free. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

[18] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[19] M. Farazmand. Multiscale analysis of accelerated gradient methods. *SIAM Journal on Optimization*, 30(3), 2020.

[20] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

[21] A. Ghosh, H. Lyu, X. Zhang, and R. Wang. Implicit regularization in heavy-ball momentum accelerated stochastic gradient descent. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZzdBhtEH9yB.

[22] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*. PMLR, 2018.

[23] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018.

[24] S. Hochreiter and J. Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994. URL https://proceedings.neurips.cc/paper_files/paper/1994/file/01882513d5fa7c329e940dda99b12147-Paper.pdf.

[25] Z. Ji and M. Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

[26] Z. Ji and M. Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[27] Z. Ji and M. Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*. PMLR, 2019.

[28] Z. Ji and M. Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[29] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

[30] T. Jules, G. Brener, T. Kachman, N. Levi, and Y. Bar-Sinai. Charting the topography of the neural network landscape with thermal-like noise. *arXiv preprint arXiv:2304.01335*, 2023. doi: 10.48550/arXiv.2304.01335. URL https://arxiv.org/abs/2304.01335.

[31] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

[32] K. Kim, S. Kotha, P. Liang, and T. Hashimoto. Pre-training under infinite compute, 2025. URL https://arxiv.org/abs/2509.14786.

[33] M. Kim, D. Li, S. X. Hu, and T. Hospedales. Fisher SAM: Information geometry and sharpness aware minimisation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022. URL https://proceedings.mlr.press/v162/kim22f.html.

[34] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] S. Kobayashi, Y. Akram, and J. von Oswald. Weight decay induces low-rank attention layers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=oDeqjIM9Sk.

[36] N. B. Kovachki and A. M. Stuart. Continuous time analysis of momentum methods. *Journal of Machine Learning Research*, 22(17), 2021.

[37] F. Kunstner, J. Chen, J. W. Lavington, and M. Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=a65YK0cqH8g.

[38] J. Kwon, J. Kim, H. Park, and I. K. Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021. URL https://proceedings.mlr.press/v139/kwon21b.html.

[39] B. Li and G. B. Giannakis. Enhancing sharpness-aware optimization through variance suppression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Sf3t6Bth4P.

[40] J. Li, A. Fang, G. Smyrnis, M. Ivgi, M. Jordan, S. Y. Gadre, H. Bansal, E. K. Guha, S. Keh, K. Arora, S. Garg, R. Xin, N. Muennighoff, R. Heckel, J. Mercat, M. F. Chen, S. Gururangan, M. Wortsman, A. Albalak, Y. Bitton, M. Nezhurina, A. K. M. Abbas, C.-Y. Hsieh, D. Ghosh, J. P. Gardner, M. Kilian, H. Zhang, R. Shao, S. M. Pratt, S. Sanyal, G. Ilharco, G. Daras, K. Marathe, A. Gokaslan, J. Zhang, K. Chandu, T. Nguyen, I. Vasiljevic, S. M. Kakade, S. Song, S. Sanghavi, F. Faghri, S. Oh, L. Zettlemoyer, K. Lo, A. El-Nouby, H. Pouransari, A. T. Toshev, S. Wang, D. Groeneveld, L. Soldaini, P. W. Koh, J. Jitsev, T. Kollar, A. Dimakis, Y. Carmon, A. Dave, L. Schmidt, and V. Shankar. Datacomp-LM: In search of the next generation of training sets for language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=CNWdWn47IE.

[41] T. Li, P. Zhou, Z. He, X. Cheng, and X. Huang. Friendly sharpness-aware minimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/CVPR52733.2024.00538.

[42] Y. Liu, S. Mai, X. Chen, C.-J. Hsieh, and Y. You. Towards efficient and scalable sharpness-aware minimization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. doi: 10.1109/CVPR52688.2022.01204.

[43] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

[44] K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[45] C. Ma, L. Wu, and W. E. A qualitative study of the dynamic behavior for adaptive gradient algorithms. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*. PMLR, 2022. URL https://proceedings.mlr.press/v145/ma22a.html.

[46] S. Malladi, K. Lyu, A. Panigrahi, and S. Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

[47] M. Marek, S. Lotfi, A. Somasundaram, A. G. Wilson, and M. Goldblum. Small batch size training for language models: When vanilla SGD works, and why gradient accumulation is wasteful. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=52Ehpe0Lu5.

[48] S. McCandlish, J. Kaplan, D. Amodei, and O. D. Team. An empirical model of large-batch training, 2018. URL https://arxiv.org/abs/1812.06162.

[49] S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

[50] T. Miyagawa. Toward equation of motion for deep neural networks: Continuous-time gradient descent and discretization error analysis. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=qq84D17BPu.

[51] M. S. Nacson, S. Gunasekar, J. Lee, N. Srebro, and D. Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In *International Conference on Machine Learning*. PMLR, 2019.

[52] M. S. Nacson, J. Lee, S. Gunasekar, P. H. P. Savarese, N. Srebro, and D. Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.

[53] M. S. Nacson, N. Srebro, and D. Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.

[54] A. Orvieto and R. M. Gower. In search of Adam's secret sauce. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=CH72XyZs4y.

[55] M. Pagliardini, P. Ablin, and D. Grangier. The adEMAMix optimizer: Better, faster, older. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jj7b3p5kLY.

[56] T. Porian, M. Wortsman, J. Jitsev, L. Schmidt, and Y. Carmon. Resolving discrepancies in compute-optimal scaling of language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=4fSSqpk1sM.

[57] Q. Qian and X. Qian. The implicit bias of adagrad on separable data. *Advances in Neural Information Processing Systems*, 32, 2019.

[58] A. Rangamani, N. H. Nguyen, A. Kumar, D. Phan, S. P. Chin, and T. D. Tran. A scale invariant measure of flatness for deep network minima. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. doi: 10.1109/ICASSP39728.2021.9413771.

[59] M. Rosca, Y. Wu, C. Qin, and B. Dherin. On a continuous time model of gradient descent dynamics and instability in deep learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=EYrRzKPinA.

[60] R. M. Schmidt, F. Schneider, and P. Hennig. Descending through a crowded valley - benchmarking deep learning optimizers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021. URL https://proceedings.mlr.press/v139/schmidt21a.html.

[61] N. Shazeer and M. Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*. PMLR, 2018.

[62] P. T. Sivaprasad, F. Mai, T. Vogels, M. Jaggi, and F. Fleuret. Optimizer benchmarking needs to account for hyperparameter tuning. In *International conference on machine learning*. PMLR, 2020.

[63] S. L. Smith, B. Dherin, D. Barrett, and S. De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=rq_Qr0c1Hyo.

[64] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1), 2018.

[65] J. M. Springer, S. Goyal, K. Wen, T. Kumar, X. Yue, S. Malladi, G. Neubig, and A. Raghunathan. Overtrained language models are harder to fine-tune. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=YW6edSufht.

[66] B. Tahmasebi, A. Soleymani, D. Bahri, S. Jegelka, and P. Jaillet. A universal class of sharpness-aware minimization algorithms. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=9Ub6nLqdMo.

[67] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[68] Y. Tsuzuku, I. Sato, and M. Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using PAC-Bayesian analysis. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020. URL https://proceedings.mlr.press/v119/tsuzuku20a.html.

[69] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*, 2024.

[70] B. Wang, Q. Meng, W. Chen, and T.-Y. Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021. URL https://proceedings.mlr.press/v139/wang21q.html.

[71] B. Wang, Q. Meng, H. Zhang, R. Sun, W. Chen, Z.-M. Ma, and T.-Y. Liu. Does momentum change the implicit regularization on separable data? *Advances in Neural Information Processing Systems*, 35, 2022.

[72] I. Watts, C. Li, S. Goyal, J. M. Springer, and A. Raghunathan. Sharpness-aware pretraining mitigates catastrophic forgetting. In *ICLR 2026 Workshop on Geometry-grounded Representation Learning and Generative Modeling*, 2026. URL https://openreview.net/forum?id=B2qTJi5s0M.

[73] K. Wen, D. Hall, T. Ma, and P. Liang. Fantastic pretraining optimizers and where to find them, 2025. URL https://arxiv.org/abs/2509.02046.

[74] L. Xiao. Rethinking conventional wisdom in machine learning: From generalization to scaling, 2025. URL https://arxiv.org/abs/2409.15156.

[75] S. Xie and Z. Li. Implicit bias of AdamW: ℓ_∞-norm constrained optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v235/xie24e.html.

[76] W. Xie, T. Pethick, and V. Cevher. SAMPa: Sharpness-aware minimization parallelized. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=IGnOktYDwV.

[77] Z. Xie, X. Wang, H. Zhang, I. Sato, and M. Sugiyama. Adaptive inertia: Disentangling the effects of adaptive learning rate and momentum. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022. URL https://proceedings.mlr.press/v162/xie22d.html.

[78] M. Yi, Q. Meng, W. Chen, Z.-m. Ma, and T.-Y. Liu. Positively scale-invariant flatness of relu neural networks. *arXiv preprint arXiv:1903.02237*, 2019. URL https://arxiv.org/abs/1903.02237.

[79] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[80] G. Zhang, C. Wang, B. Xu, and R. Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1lz-3Rct7.

[81] H. Zhang, D. Morwani, N. Vyas, J. Wu, D. Zou, U. Ghai, D. Foster, and S. M. Kakade. How does critical batch size scale in pre-training? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=JCiF03qnmi.

[82] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[83] R. Zhao, D. Morwani, D. Brandfonbrener, N. Vyas, and S. M. Kakade. Deconstructing what makes a good optimizer for autoregressive language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zfeso8ceqr.

[84] Y. Zheng, R. Zhang, and Y. Mao. Regularizing neural networks via adversarial model perturbation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. doi: 10.1109/CVPR46437.2021.00806.

[85] P. Zhou, J. Feng, C. Ma, C. Xiong, S. C. H. Hoi, and W. E. Towards theoretically understanding why sgd generalizes better than adam in deep learning. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/f3f27a324736617f20abbf2ffd806f6d-Paper.pdf.

[86] P. Zhou, X. Xie, Z. Lin, and S. Yan. Towards understanding convergence and generalization of AdamW. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9), 2024. doi: 10.1109/TPAMI.2024.3382294.

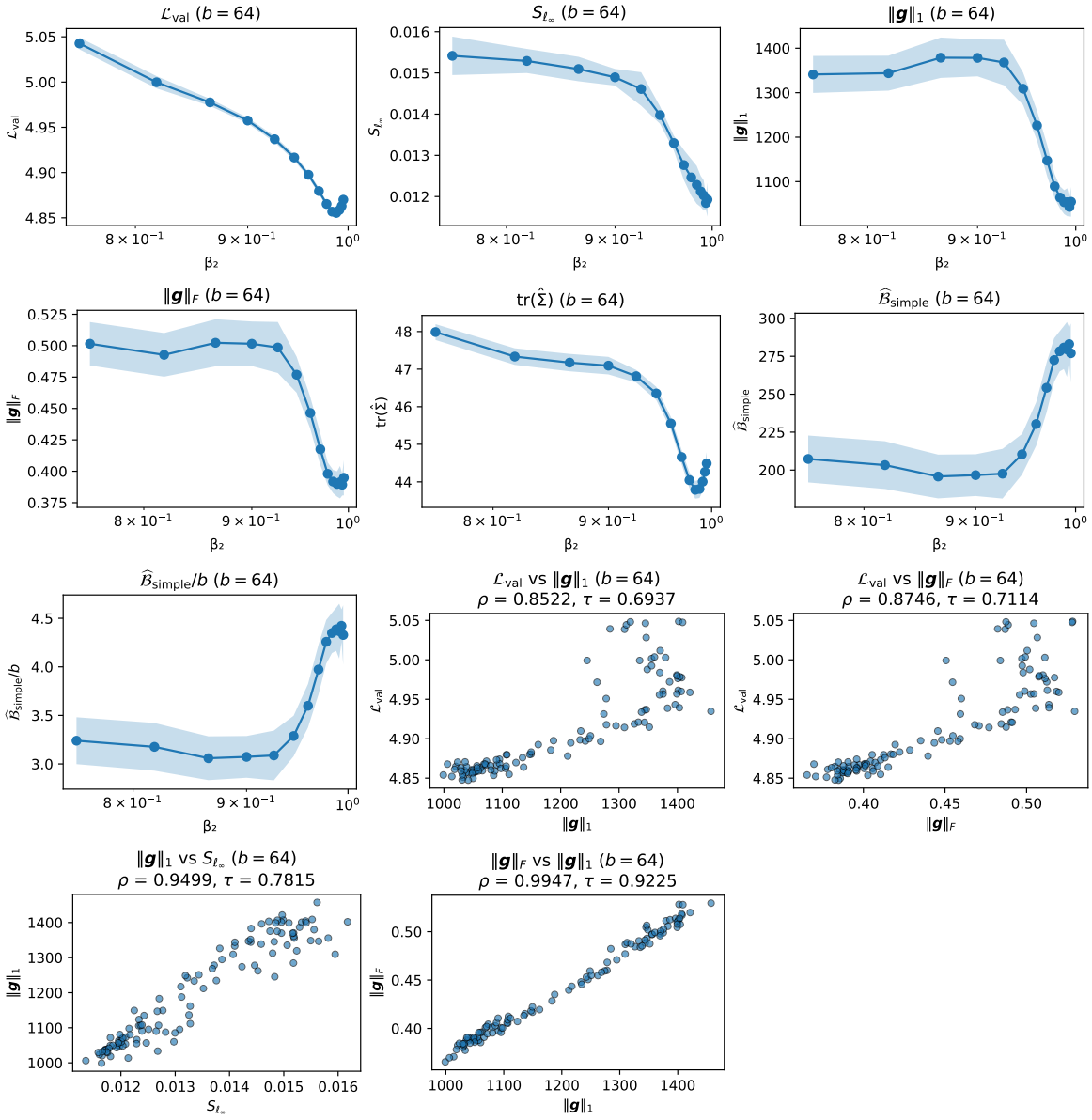
[87] Z. Zhuang, M. Liu, A. Cutkosky, and F. Orabona. Understanding adamw through proximal methods and scale-freeness, 2022. URL https://openreview.net/forum?id=GU11Lbci5J.

A Further Evidence and Experiment Details

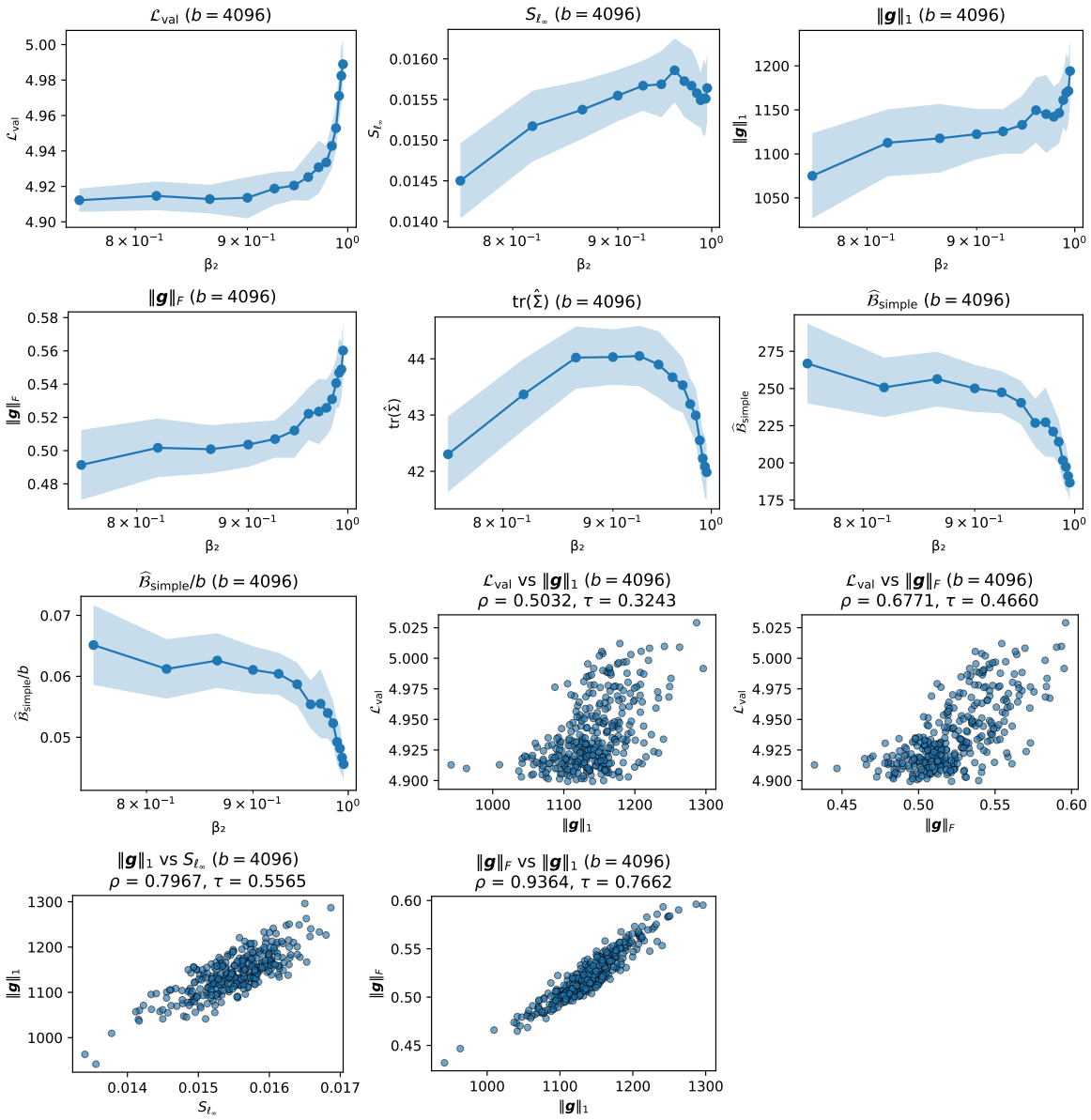
A.1 Transformer-XL on WikiText-2: β_2 Sweep with $\beta_1 = 0.9$ Fixed

Transformer-XL [14] with about 55 M parameters is trained on WikiText-2 [49] following Kunstner et al. [37] until after overfitting on the training set (when the validation loss starts rising). Adam is used without weight decay, $\epsilon = 10^{-6}$. The learning rate is constant 10^{-4}. For $b = 64$, 16 iterations completed and the sharpness metrics are plotted at epoch 9 (median overfitting epoch 9, mean 8.5); for $b = 4096$, 32 iterations completed and the sharpness metrics are plotted at epoch 100 (median overfitting epoch 103, mean 99.2). We plot the validation loss, ℓ_∞ sharpness calculated by projected gradient ascent and noise metrics at the last step, along with scatterplots showing correlations; ρ in the titles denotes the Pearson correlation coefficient and τ denotes Kendall's tau. Validation loss does not use EMA but other metrics use EMA with parameter $\beta = 0.99$.

Sharpness Metrics at Batch Size 64



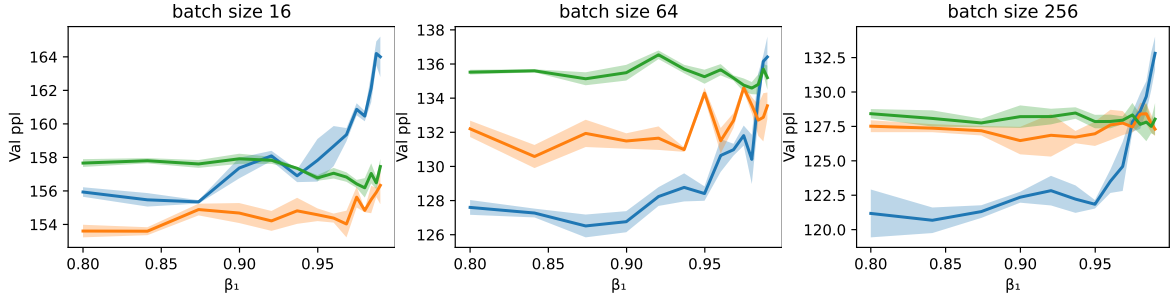
Sharpness Metrics at Batch Size 4096

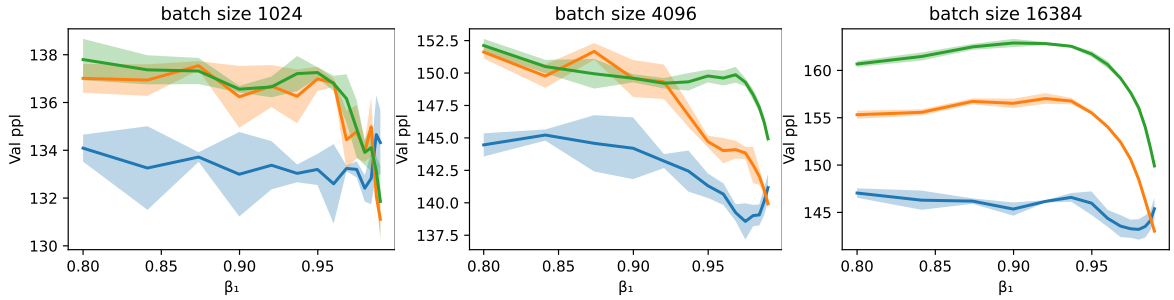


A.2 Transformer-XL on WikiText-2: β_1 Sweep with $\beta_2 = 0.999$ Fixed

A.2.1 Minimal Validation Perplexity at a Large Set of Batch Sizes

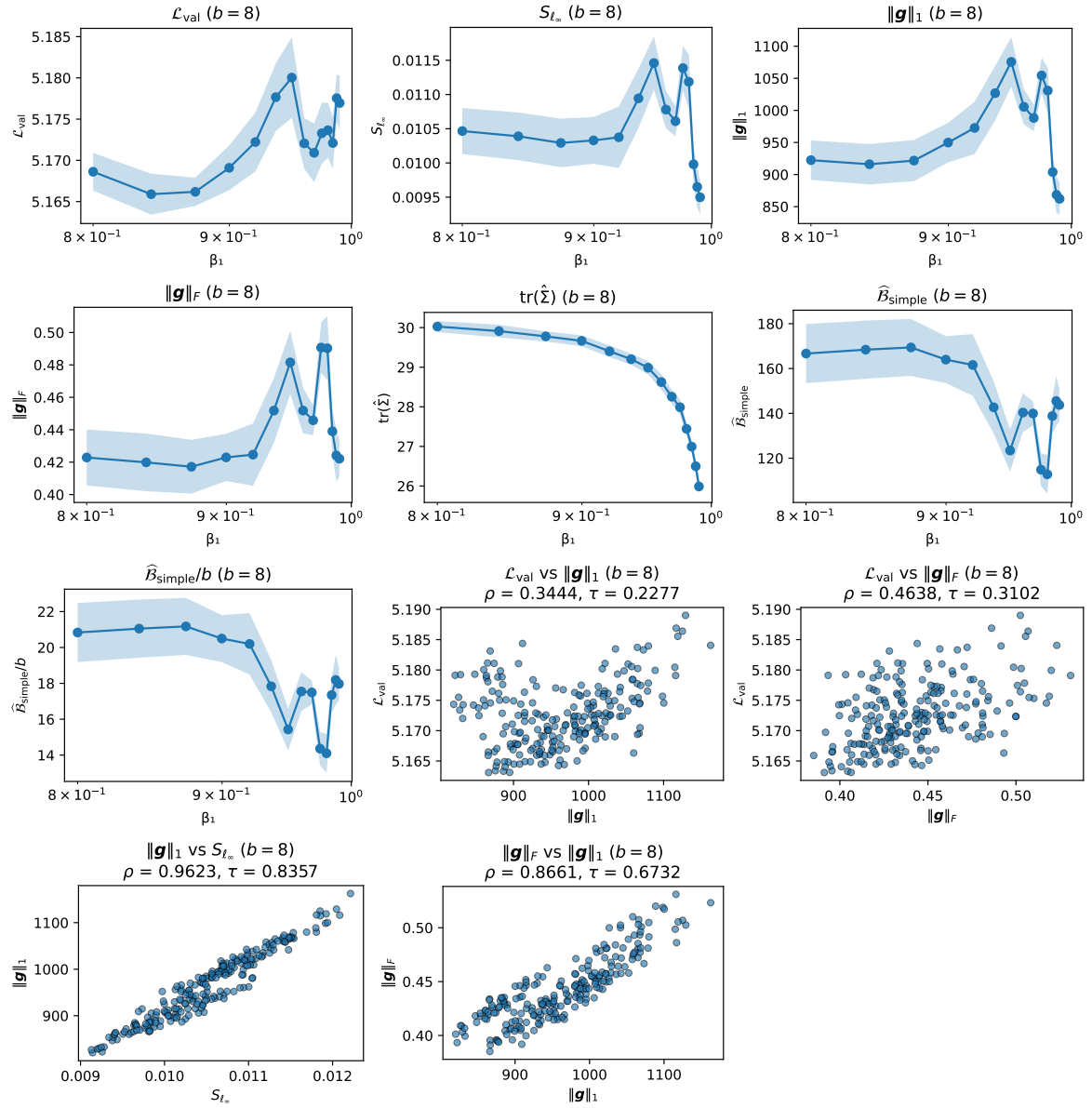
We plot the minimal validation perplexity achieved (by definition, it is exactly at the point of overfitting), with learning rates $\{10^{-3.5}, 10^{-4}, 10^{-4.5}\}$ and different batch sizes. The monotonicity trends as a function of β_1 largely revert as the batch size increases.



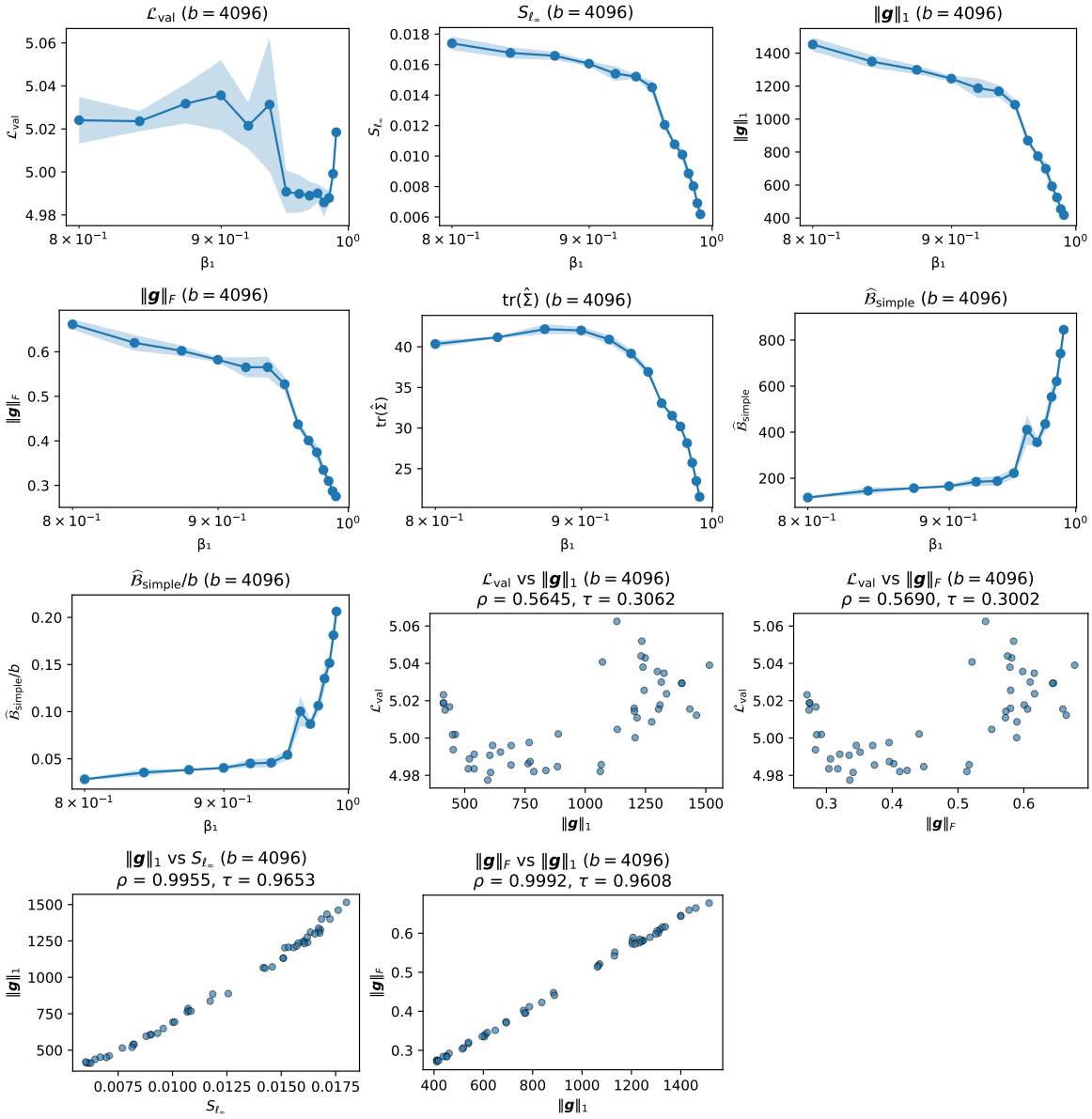


Similarly to the β_2 sweep, we also plot below the sharpness metrics.

A.2.2 Sharpness Metrics at Batch Size 8

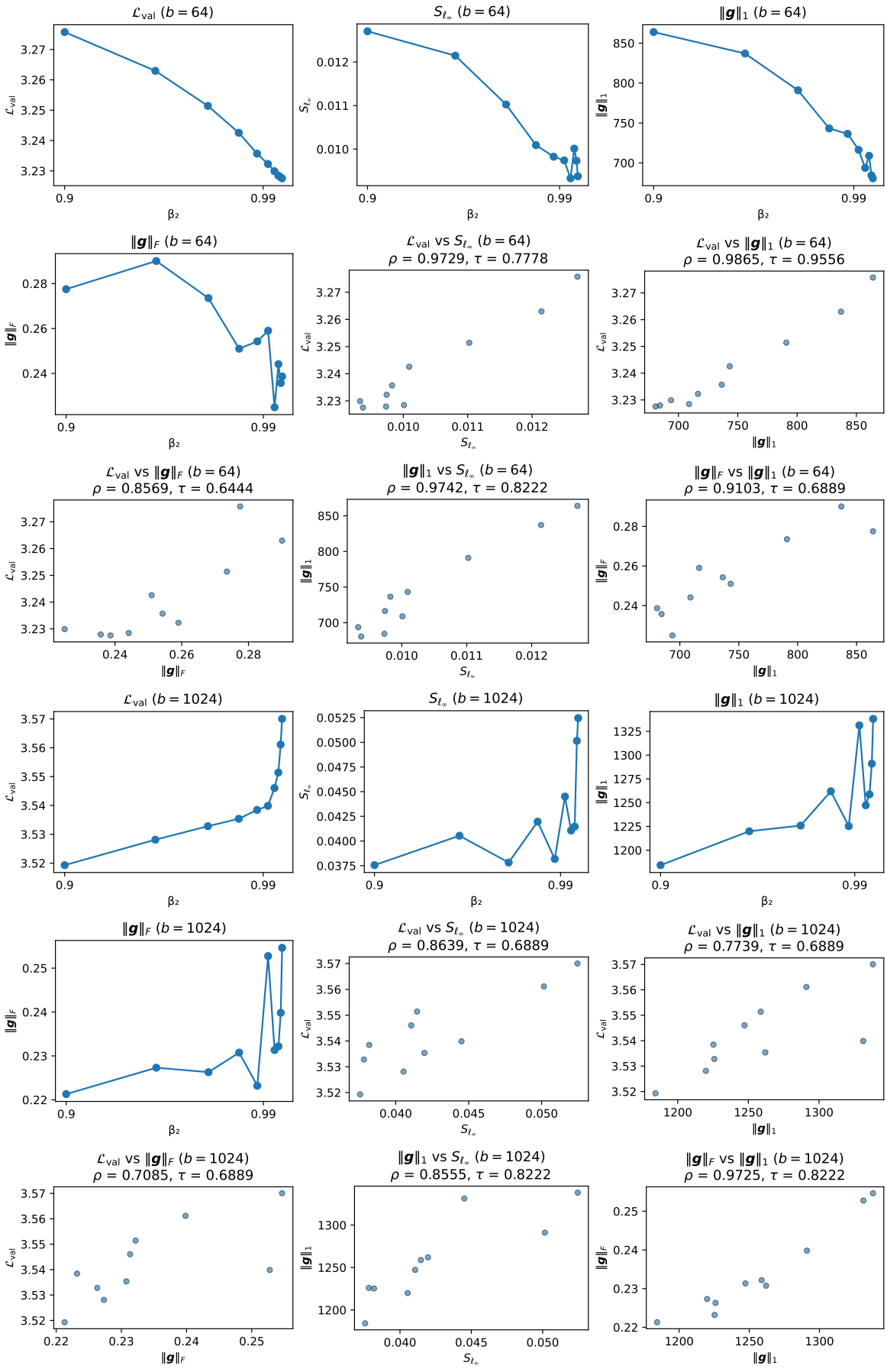


A.2.3 Sharpness Metrics at Batch Size 4096



A.3 Llama3 on DCLM

Llama3 with 12 layers and 12 heads, head dimension 64, sequence length 4096 is trained on DCLM for 6 144 000 sequences (25 B tokens, or about 90 tokens per parameter). No weight tying is used, and the total number of parameters is around 280 M. The dataset is tokenized with the Llama 3.1 tokenizer. We isolate the first 2^{24} documents of DCLM as a validation set, and choose the first 1024 sequences (4 M tokens) for calculating the validation loss and other metrics. AdamW is used with $\epsilon = 10^{-8}$ and decoupled weight decay $\lambda = 0.1$ (PyTorch parametrization) applied to all but the normalization parameters. The number of training steps is chosen to make a full pass over the sequences (96 000 steps for $b = 64$ and 6 000 steps for $b = 1024$). We plot validation loss and sharpness metrics at the last step, along with scatterplots showing correlations (ρ in the title denotes the Pearson correlation coefficient, τ denotes Kendall's tau).


$\mathcal{L}_{\text{val}}$ ($b = 64$)
S_{ℓ_∞} ($b = 64$)
$\|\boldsymbol{g}\|_1$ ($b = 64$)
$\|\boldsymbol{g}\|_F$ ($b = 64$)
$\mathcal{L}_{\text{val}}$ vs S_{ℓ_∞} ($b = 64$)
ρ = 0.9729, τ = 0.7778
$\mathcal{L}_{\text{val}}$ vs $\|\boldsymbol{g}\|_1$ ($b = 64$)
ρ = 0.9865, τ = 0.9556
$\mathcal{L}_{\text{val}}$ vs $\|\boldsymbol{g}\|_F$ ($b = 64$)
ρ = 0.8569, τ = 0.6444
$\|\boldsymbol{g}\|_1$ vs S_{ℓ_∞} ($b = 64$)
ρ = 0.9742, τ = 0.8222
$\|\boldsymbol{g}\|_F$ vs $\|\boldsymbol{g}\|_1$ ($b = 64$)
ρ = 0.9103, τ = 0.6889
$\mathcal{L}_{\text{val}}$ ($b = 1024$)
S_{ℓ_∞} ($b = 1024$)
$\|\boldsymbol{g}\|_1$ ($b = 1024$)
$\|\boldsymbol{g}\|_F$ ($b = 1024$)
$\mathcal{L}_{\text{val}}$ vs S_{ℓ_∞} ($b = 1024$)
ρ = 0.8639, τ = 0.6889
$\mathcal{L}_{\text{val}}$ vs $\|\boldsymbol{g}\|_1$ ($b = 1024$)
ρ = 0.7739, τ = 0.6889
$\mathcal{L}_{\text{val}}$ vs $\|\boldsymbol{g}\|_F$ ($b = 1024$)
ρ = 0.7085, τ = 0.6889
$\|\boldsymbol{g}\|_1$ vs S_{ℓ_∞} ($b = 1024$)
ρ = 0.8555, τ = 0.8222
$\|\boldsymbol{g}\|_F$ vs $\|\boldsymbol{g}\|_1$ ($b = 1024$)
ρ = 0.9725, τ = 0.8222
β_2
$\mathcal{L}_{\text{val}}$
S_{ℓ_∞}
$\|\boldsymbol{g}\|_1$
$\|\boldsymbol{g}\|_F$

B Main Theorems

We start with the full statement of the memory removal step (Section 3.1), from which technical expressions were omitted in the main part of this article.

Theorem B.1 (Memory removal)**.** *Let $\Theta \subset \mathbb{R}^{\dim \boldsymbol{\theta}}$ be an open convex domain of parameters $\boldsymbol{\theta}$ of interest, and assume $\ell_r(\cdot) \in \mathcal{C}^3(\Theta; \mathbb{R})$ with*

$$\sup_{\boldsymbol{\theta}\in\Theta} \sup_{N} \max_{1\le r\le N} \max_{1\le s\le 3} \|\nabla^s \ell_r(\boldsymbol{\theta})\| < \infty,$$

where $\|\cdot\|$ is the corresponding operator norm. Let the iteration $\{\boldsymbol{\theta}_t\}_{t=0}^{\infty}$ be given by Adam, Definition 3.1, and the iteration $\{\tilde{\boldsymbol{\theta}}_t\}_{t=0}^{\infty}$ be given by

$$\begin{aligned}
&\tilde{\boldsymbol{\theta}}_{t+1} = \tilde{\boldsymbol{\theta}}_t - \eta\, \boldsymbol{Main}_t(\tilde{\boldsymbol{\theta}}_t) - \eta^2\, \boldsymbol{Corr}_t(\tilde{\boldsymbol{\theta}}_t), \quad \tilde{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0,\\
\text{with}\quad & Main_{t,j}(\boldsymbol{\theta}) := \frac{\sum_{k=0}^{t} \mu_{t,k} \partial_j \mathcal{L}_k(\boldsymbol{\theta})}{\sqrt{\sum_{k=0}^{t} \nu_{t,k} |\partial_j \mathcal{L}_k(\boldsymbol{\theta})|^2 + \epsilon}}, \qquad (10)\\
& Corr_{t,j}(\boldsymbol{\theta}) := \frac{L_{t,j}(\boldsymbol{\theta})}{R_{t,j}(\boldsymbol{\theta})} - \frac{M_{t,j}(\boldsymbol{\theta}) P_{t,j}(\boldsymbol{\theta})}{R_{t,j}(\boldsymbol{\theta})^3}, \qquad (11)\\
& M_{t,j}(\boldsymbol{\theta}) := \sum_{k=0}^{t} \mu_{t,k} \partial_j \mathcal{L}_k(\boldsymbol{\theta}),\\
& R_{t,j}(\boldsymbol{\theta}) := \sqrt{\sum_{k=0}^{t} \nu_{t,k} |\partial_j \mathcal{L}_k(\boldsymbol{\theta})|^2 + \epsilon},\\
& L_{t,j}(\boldsymbol{\theta}) := \sum_{k=0}^{t-1} \mu_{t,k} \sum_{i=1}^{\dim \boldsymbol{\theta}} \partial_{ij} \mathcal{L}_k(\boldsymbol{\theta}) \sum_{l=k}^{t-1} \frac{M_{l,i}(\boldsymbol{\theta})}{R_{l,i}(\boldsymbol{\theta})},\\
& P_{t,j}(\boldsymbol{\theta}) := \sum_{k=0}^{t-1} \nu_{t,k} \partial_j \mathcal{L}_k(\boldsymbol{\theta}) \sum_{i=1}^{\dim \boldsymbol{\theta}} \partial_{ij} \mathcal{L}_k(\boldsymbol{\theta}) \sum_{l=k}^{t-1} \frac{M_{l,i}(\boldsymbol{\theta})}{R_{l,i}(\boldsymbol{\theta})}.
\end{aligned}$$

Then, for any constant "physical time" horizon $T > 0$, the following global error bound holds:

$$\max_{t\in[0:\lfloor T/\eta \rfloor]} \left\|\boldsymbol{\theta}_t - \tilde{\boldsymbol{\theta}}_t\right\|_\infty \le C\eta^2$$

for some constant $C = C_T$ not depending on η.

Next, we state the full version of the mini-batch noise expansions.

Theorem B.2 (Mini-batch noise expansion of the memoryless dynamics)**.** *In the setting of Theorem B.1, for every $j \in [1 : \dim \boldsymbol{\theta}]$,*

$$\begin{aligned}
\sqrt{g_j^2 + \epsilon}\, \mathbb{E}_\pi Corr_{n,j} = {} & FB_j^{(n,\epsilon)} + MBN_{1,j}^{(n,\epsilon)} + MBN_{2,j}^{(n,\epsilon)}\\
& + MBN_{3,j}^{(n,\epsilon)} + MBN_{4,j}^{(n,\epsilon)} + MBN_{5,j}^{(n,\epsilon)}\\
& + O(d^3) + o_n(b^{-1}),
\end{aligned} \qquad (12)$$

where

$$FB_j^{(n,\epsilon)} := \left(\sum_{k=0}^{n-1} \mu_{n,k}(n-k) - \frac{g_j^2}{g_j^2 + \epsilon} \sum_{k=0}^{n-1} \nu_{n,k}(n-k) \right) \partial_j \|\boldsymbol{g}\|_{1,\epsilon},$$

and the five mini-batch-noise corrections are given by

$$\begin{aligned}
MBN_{1,j}^{(n,\epsilon)} &:= \frac{n}{(n+1)b-1}\, \partial_j \|\boldsymbol{g}\|_{1,\epsilon}\, \Sigma_{jj}\, A_j^{(n,\epsilon)},\\
MBN_{2,j}^{(n,\epsilon)} &:= \frac{n}{(n+1)b-1} \sum_i \frac{\partial_i g_j g_i}{(g_i^2 + \epsilon)^{5/2}} \Sigma_{ii}\, B_{i,j}^{(n,\epsilon)},
\end{aligned}$$

$$
\begin{aligned}
MBN_{3,j}^{(n,\epsilon)} &:= \frac{n}{2((n+1)b-1)}\frac{g_j}{g_j^2+\epsilon} \\
&\quad\times\left(-2\sum_{k=0}^{n-1}(n-k)\mu_{n,k}\nu_{n,k} - \sum_{k=0}^{n-1}(n-k)\nu_{n,k} + 3\frac{g_j^2}{g_j^2+\epsilon}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}^2\right) \\
&\quad\times\sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}}\,\partial_i\Sigma_{jj}, \\
MBN_{4,j}^{(n,\epsilon)} &:= \frac{n}{2((n+1)b-1)}\sum_i \frac{1}{\sqrt{g_i^2+\epsilon}} D_{i,j}^{(n,\epsilon)}\,\partial_j\Sigma_{ii}, \\
MBN_{5,j}^{(n,\epsilon)} &:= \frac{n}{(n+1)b-1}\frac{g_j}{g_j^2+\epsilon}\sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}} E_{i,j}^{(n,\epsilon)}\,\Sigma_{ij}.
\end{aligned}
$$

Here

$$
\begin{aligned}
A_j^{(n,\epsilon)} &:= \frac{1}{2(g_j^2+\epsilon)^2}\sum_{r=0}^{n}\big(3g_j^2\nu_{n,r}^2 - (g_j^2+\epsilon)\nu_{n,r}\big)\sum_{k=0}^{n-1}\mu_{n,k}(n-k) \\
&\quad - \frac{g_j^2}{(g_j^2+\epsilon)^3}\sum_{r=0}^{n-1}(n-r)\nu_{n,r}\sum_{k=0}^{n}\big(4g_j^2\nu_{n,k}^2 - (g_j^2+\epsilon)\nu_{n,k} - 2(g_j^2+\epsilon)\mu_{n,k}\nu_{n,k}\big) \\
&\quad - \frac{1}{(g_j^2+\epsilon)^2}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\big((g_j^2+\epsilon)\mu_{n,k} - 2g_j^2\nu_{n,k}\big) \\
&\quad + \frac{g_j^2}{(g_j^2+\epsilon)^3}\sum_{p=0}^{n-1}(n-p)\nu_{n,p}\sum_{k=0}^{n}\nu_{n,k}\big((g_j^2+\epsilon)\mu_{n,k} - 2g_j^2\nu_{n,k}\big) \\
&\quad - \frac{g_j^2}{2(g_j^2+\epsilon)^3}\sum_{r=0}^{n-1}(n-r)\nu_{n,r}\sum_{k=0}^{n}\big(3g_j^2\nu_{n,k}^2 - (g_j^2+\epsilon)\nu_{n,k}\big) \\
&\quad + \frac{g_j^2}{(g_j^2+\epsilon)^2}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}^2, \\
B_{i,j}^{(n,\epsilon)} &:= \frac{1}{2}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\big(3g_i^2\nu_{l,p}^2 - (g_i^2+\epsilon)\nu_{l,p} - 2(g_i^2+\epsilon)\mu_{l,p}\nu_{l,p}\big) \\
&\quad - \frac{g_j^2}{2(g_j^2+\epsilon)}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\big(3g_i^2\nu_{l,p}^2 - (g_i^2+\epsilon)\nu_{l,p} - 2(g_i^2+\epsilon)\mu_{l,p}\nu_{l,p}\big), \\
D_{i,j}^{(n,\epsilon)} &:= \sum_{l=0}^{n-1}\sum_{k=0}^{l}\mu_{n,k}\left[\mu_{l,k} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,k}\right] - \frac{g_j^2}{g_j^2+\epsilon}\sum_{l=0}^{n-1}\sum_{k=0}^{l}\nu_{n,k}\left[\mu_{l,k} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,k}\right], \\
E_{i,j}^{(n,\epsilon)} &:= -\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\nu_{n,p}\left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right] \\
&\quad - \frac{1}{g_j^2+\epsilon}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\big((g_j^2+\epsilon)\mu_{n,p} - 2g_j^2\nu_{n,p}\big) \\
&\quad + \frac{g_j^2}{g_j^2+\epsilon}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\nu_{n,p} \\
&\quad - \sum_{l=0}^{n-1}\sum_{k=0}^{l}\nu_{n,k}\left[\mu_{l,k} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,k}\right].
\end{aligned}
$$

The following theorem provides limits of the above expressions which is useful for interpreting them.

Theorem B.3 (Limits of full-batch and mini-batch corrections)**.**

(a) The finite-n quantities in Theorem B.2 satisfy, as $n \to \infty$,

$$FB_j^{(n,\epsilon)} = FB_j^{(\infty,\epsilon)} + o_n(1), \qquad MBN_{r,j}^{(n,\epsilon)} = MBN_{r,j}^{(\infty,\epsilon)} + o_n(b^{-1}), \qquad r \in [1\!:\!5],$$

where

$$\begin{aligned}
FB_j^{(\infty,\epsilon)} &:= \left(\frac{\beta_1}{1-\beta_1} - \frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2}{1-\beta_2}\right)\partial_j \|\boldsymbol{g}\|_{1,\epsilon},\\
MBN_{1,j}^{(\infty,\epsilon)} &:= \frac{1}{b}\,\partial_j \|\boldsymbol{g}\|_{1,\epsilon}\,\Sigma_{jj}\,A_j^{(\infty,\epsilon)},\\
MBN_{2,j}^{(\infty,\epsilon)} &:= \frac{1}{b}\sum_i \frac{\partial_i g_j g_i}{(g_i^2+\epsilon)^{5/2}}\Sigma_{ii}\,B_{i,j}^{(\infty,\epsilon)},\\
MBN_{3,j}^{(\infty,\epsilon)} &:= \frac{1}{2b}\frac{g_j}{g_j^2+\epsilon}\left(-2\frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2} - \frac{\beta_2}{1-\beta_2} + 3\frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2^2}{(1+\beta_2)^2}\right)\\
&\quad\times\sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}}\,\partial_i\Sigma_{jj},\\
MBN_{4,j}^{(\infty,\epsilon)} &:= \frac{1}{2b}\sum_i \frac{1}{\sqrt{g_i^2+\epsilon}}D_{i,j}^{(\infty,\epsilon)}\,\partial_j\Sigma_{ii},\\
MBN_{5,j}^{(\infty,\epsilon)} &:= \frac{1}{b}\frac{g_j}{g_j^2+\epsilon}\sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}E_{i,j}^{(\infty,\epsilon)}\,\Sigma_{ij}.
\end{aligned}$$

The limiting coefficient functions are

$$\begin{aligned}
A_j^{(\infty,\epsilon)} &:= \frac{\beta_1}{2(1-\beta_1)(g_j^2+\epsilon)}\left(3\frac{g_j^2}{g_j^2+\epsilon}\frac{1-\beta_2}{1+\beta_2} - 1\right)\\
&\quad - \frac{g_j^2}{(g_j^2+\epsilon)^2}\frac{\beta_2}{1-\beta_2}\left(4\frac{g_j^2}{g_j^2+\epsilon}\frac{1-\beta_2}{1+\beta_2} - 1 - \frac{2(1-\beta_1)(1-\beta_2)}{1-\beta_1\beta_2}\right)\\
&\quad - \frac{1}{g_j^2+\epsilon}\left(\frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2} - 2\frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2^2}{(1+\beta_2)^2}\right)\\
&\quad + \frac{g_j^2}{(g_j^2+\epsilon)^2}\frac{\beta_2}{1-\beta_2}\left(\frac{(1-\beta_1)(1-\beta_2)}{1-\beta_1\beta_2} - 2\frac{g_j^2}{g_j^2+\epsilon}\frac{1-\beta_2}{1+\beta_2}\right)\\
&\quad - \frac{g_j^2}{2(g_j^2+\epsilon)^2}\frac{\beta_2}{1-\beta_2}\left(3\frac{g_j^2}{g_j^2+\epsilon}\frac{1-\beta_2}{1+\beta_2} - 1\right) + \frac{g_j^2}{(g_j^2+\epsilon)^2}\frac{\beta_2^2}{(1+\beta_2)^2},\\
B_{i,j}^{(\infty,\epsilon)} &:= \frac{g_i^2+\epsilon}{2}\left(3\frac{g_i^2}{g_i^2+\epsilon}\frac{1-\beta_2}{1+\beta_2} - 1 - \frac{2(1-\beta_1)(1-\beta_2)}{1-\beta_1\beta_2}\right)\\
&\quad\times\left(\frac{\beta_1}{1-\beta_1} - \frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2}{1-\beta_2}\right),\\
D_{i,j}^{(\infty,\epsilon)} &:= \frac{\beta_1}{1+\beta_1} - \frac{g_i^2}{g_i^2+\epsilon}\frac{\beta_1(1-\beta_2)}{1-\beta_1\beta_2} - \frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2(1-\beta_1)}{1-\beta_1\beta_2}\\
&\quad + \frac{g_i^2}{g_i^2+\epsilon}\frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2}{1+\beta_2},\\
E_{i,j}^{(\infty,\epsilon)} &:= -\frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2} + \frac{g_i^2}{g_i^2+\epsilon}\frac{\beta_1\beta_2(1-\beta_2)}{(1+\beta_2)(1-\beta_1\beta_2)}\\
&\quad - \frac{\beta_1\beta_2(1-\beta_1)}{(1+\beta_1)(1-\beta_1\beta_2)} + \frac{g_i^2}{g_i^2+\epsilon}\frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2}
\end{aligned}$$

$$+3\frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2^2(1-\beta_1)}{(1+\beta_2)(1-\beta_1\beta_2)}-3\frac{g_i^2}{g_i^2+\epsilon}\frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2^2}{(1+\beta_2)^2}$$
$$-\frac{\beta_2(1-\beta_1)}{1-\beta_1\beta_2}+\frac{g_i^2}{g_i^2+\epsilon}\frac{\beta_2}{1+\beta_2}.$$

(b) Further, the limits of the full-batch and mini-batch noise expansions as $\epsilon \to 0$

$$FB_j^{(\infty,\epsilon)} = FB_j + o_\epsilon(1), \quad MBN_{r,j}^{(\infty,\epsilon)} = MBN_{r,j} + o_\epsilon(1), \quad r \in [1:5],$$

are given by Eqs. (7) and (8) with

$$\begin{aligned}
C_1(\beta_1,\beta_2) &:= \frac{1-\beta_1^2}{\beta_1(1-\beta_1\beta_2)} + \frac{(1-\beta_1)^2}{\beta_1(1-\beta_1\beta_2)^2} + \frac{3(1+\beta_1)}{2(1-\beta_1)(1+\beta_2)} \\
&\quad - \frac{2}{\beta_1(1-\beta_1)} + \frac{3}{2-2\beta_2} + \frac{3}{(1+\beta_2)^2} - 2, \\
C_2(\beta_1,\beta_2) &:= \frac{(\beta_1-\beta_2)(\beta_1\beta_2^2-\beta_1\beta_2+\beta_1+\beta_2^2-2\beta_2)}{(1-\beta_1)(1-\beta_2)(1+\beta_2)(1-\beta_1\beta_2)}, \\
C_3(\beta_1,\beta_2) &:= \frac{1}{2(1-\beta_2)(1+\beta_2)^2(1-\beta_1\beta_2)^2}\big\{-2\beta_1(\beta_1+1)\beta_2^5 + (\beta_1^2+8\beta_1)\beta_2^4 \\
&\quad + (2\beta_1-5\beta_1^2-4)\beta_2^3 + (2\beta_1+1)\beta_2^2 + (2\beta_1^2-2\beta_1-1)\beta_2\big\}, \\
C_4(\beta_1,\beta_2) &:= -\frac{(\beta_1-\beta_2)^2}{2(1+\beta_1)(1+\beta_2)(1-\beta_1\beta_2)}, \\
C_5(\beta_1,\beta_2) &:= \frac{\beta_2(\beta_2-\beta_1)(2\beta_2-3\beta_1-1)}{(1+\beta_1)(1+\beta_2)^2(1-\beta_1\beta_2)}.
\end{aligned} \tag{13}$$

C Proof of Theorem B.1

This result is taken from Cattaneo and Shigida [10]. Specifically, it is a special case of the following general theorem, a reformulation of their Corollary 3.3.

Theorem C.1 (General memory removal theorem)**.** *Let Θ be an open convex domain in $\mathbb{R}^{\dim \boldsymbol{\theta}}$ and $\boldsymbol{F}_t \in C^2(\Theta^{t+1};\mathbb{R}^d)$ be a family of functions, such that for any $t \in \mathbb{Z}_{\geq 0}$, $k_1,k_2 \in [0:t]$, $r,i,j \in [1:\dim\boldsymbol{\theta}]$,*

$$|F_{t,r}| \leq \gamma_{-1}, \quad \left|\frac{\partial F_{t,r}}{\partial \theta_{t-k_1,i}}\right| \leq \gamma_{k_1}, \quad \left|\frac{\partial^2 F_{t,r}}{\partial \theta_{t-k_1,i}\partial\theta_{t-k_2,j}}\right| \leq \gamma_{k_1,k_2},$$

where γ_{-1}, γ_{k_1} and γ_{k_1,k_2} are families of positive reals (not depending on t) satisfying $\sum_{k_1=1}^\infty \gamma_{k_1}k_1^2 + \sum_{k_1,k_2=1}^\infty \gamma_{k_1,k_2}k_1k_2 < \infty$ (sufficiently fast decay of memory). Let $T \geq 0$ be a fixed "physical time" horizon. Then iterations $\{\boldsymbol{\theta}_t\}_{t=0}^\infty$ and $\{\tilde{\boldsymbol{\theta}}_t\}_{t=0}^\infty$, given in Eqs. (1), (2) and (4) with the same initial condition $\tilde{\boldsymbol{\theta}}_0 = \boldsymbol{\theta}_0$, satisfy

$$\max_{t\in[0:\lfloor T/\eta\rfloor]} \left\|\boldsymbol{\theta}_t - \tilde{\boldsymbol{\theta}}_t\right\|_\infty \leq C\eta^2$$

for some constant C depending on T but not depending on η.

The proof of Theorem B.1 is a direct application of Theorem C.1, with the function $\boldsymbol{F}_t$ given by

$$F_{t,j}(\boldsymbol{\theta}_t,\ldots,\boldsymbol{\theta}_0) := \frac{\sum_{k=0}^t \mu_{t,k}\partial_j\mathcal{L}_k(\boldsymbol{\theta}_k)}{\sqrt{\sum_{k=0}^t \nu_{t,k}|\partial_j\mathcal{L}_k(\boldsymbol{\theta}_k)|^2+\epsilon}}.$$

We check the assumptions below.

By the boundedness assumption on the per-sample losses, there exist constants $G,H,K<\infty$ such that, uniformly in the batch index and in $\boldsymbol{\theta}\in\Theta$,

$$\|\nabla\mathcal{L}_k(\boldsymbol{\theta})\| \leq G, \qquad \|\nabla^2\mathcal{L}_k(\boldsymbol{\theta})\| \leq H, \qquad \|\nabla^3\mathcal{L}_k(\boldsymbol{\theta})\| \leq K.$$

Since $\sum_{k=0}^{t}\mu_{t,k}=1$ and $\sum_{k=0}^{t}\nu_{t,k}=1$, we have a bound $|F_{t,j}(\boldsymbol{\theta}_t,\dots,\boldsymbol{\theta}_0)|\le G/\sqrt{\epsilon}$. Hence the zeroth-order bound in Theorem C.1 holds with, for example, $\gamma_{-1}:=G/\sqrt{\epsilon}$.

Now fix $a\in[0:t]$. Differentiating $F_{t,j}$ with respect to the coordinate $\theta_{a,i}$ gives

$$\frac{\partial F_{t,j}}{\partial\theta_{a,i}}=\frac{\mu_{t,a}\partial_{ij}\mathcal{L}_a(\boldsymbol{\theta}_a)}{\left(\sum_{k=0}^{t}\nu_{t,k}|\partial_j\mathcal{L}_k(\boldsymbol{\theta}_k)|^2+\epsilon\right)^{1/2}}-\sum_{k=0}^{t}\mu_{t,k}\partial_j\mathcal{L}_k(\boldsymbol{\theta}_k)\frac{\nu_{t,a}\partial_j\mathcal{L}_a(\boldsymbol{\theta}_a)\partial_{ij}\mathcal{L}_a(\boldsymbol{\theta}_a)}{\left(\sum_{k=0}^{t}\nu_{t,k}|\partial_j\mathcal{L}_k(\boldsymbol{\theta}_k)|^2+\epsilon\right)^{3/2}}.$$

Using the bounds above,

$$\left|\frac{\partial F_{t,j}}{\partial\theta_{a,i}}\right|\le C_\epsilon(\mu_{t,a}+\nu_{t,a})$$

for a constant C_ϵ depending only on G,H,ϵ, but not on t or a. If $a=t-q$, then

$$\mu_{t,t-q}=\frac{(1-\beta_1)\beta_1^q}{1-\beta_1^{t+1}}\le\beta_1^q,\qquad\nu_{t,t-q}=\frac{(1-\beta_2)\beta_2^q}{1-\beta_2^{t+1}}\le\beta_2^q.$$

Thus

$$\left|\frac{\partial F_{t,j}}{\partial\theta_{t-q,i}}\right|\le C_\epsilon(\beta_1^q+\beta_2^q).$$

So we may take

$$\gamma_q:=C_\epsilon(\beta_1^q+\beta_2^q).$$

Since $0<\beta_1,\beta_2<1$,

$$\sum_{q=1}^{\infty}q^2\gamma_q<\infty.$$

Similarly, differentiating once more, every second derivative of $F_{t,j}$ is a finite linear combination of terms of the following schematic forms:

$$\mathbf{1}_{\{a=b\}}\mu_{t,a},\qquad\mathbf{1}_{\{a=b\}}\nu_{t,a},\qquad\mu_{t,a}\nu_{t,b},\qquad\nu_{t,a}\mu_{t,b},\qquad\nu_{t,a}\nu_{t,b},$$

multiplied by bounded derivatives of the losses and by powers of $\left(\sum_{k=0}^{t}\nu_{t,k}|\partial_j\mathcal{L}_k(\boldsymbol{\theta}_k)|^2+\epsilon\right)^{-1/2}$. The latter are uniformly bounded because $R_{t,j}\ge\sqrt{\epsilon}$. Therefore, if $a=t-q_1$ and $b=t-q_2$, then

$$\left|\frac{\partial^2 F_{t,j}}{\partial\theta_{t-q_1,i}\partial\theta_{t-q_2,r}}\right|\le C_\epsilon\left(\mathbf{1}_{\{q_1=q_2\}}(\beta_1^{q_1}+\beta_2^{q_1})+(\beta_1^{q_1}+\beta_2^{q_1})(\beta_1^{q_2}+\beta_2^{q_2})\right).$$

Thus we may choose

$$\gamma_{q_1,q_2}:=C_\epsilon\left(\mathbf{1}_{\{q_1=q_2\}}(\beta_1^{q_1}+\beta_2^{q_1})+(\beta_1^{q_1}+\beta_2^{q_1})(\beta_1^{q_2}+\beta_2^{q_2})\right).$$

Then

$$\sum_{q_1,q_2=1}^{\infty}\gamma_{q_1,q_2}q_1q_2<\infty,$$

again because $0<\beta_1,\beta_2<1$. Hence all hypotheses of Theorem C.1 are satisfied.

The main and correction terms from Eqs. (10) and (11) are obtained by directly using Eq. (4) (see also Cattaneo and Shigida [10]).

D Proof of Theorem B.2

The plan is to first expand $\mathrm{Corr}_{n,j}(\boldsymbol{\theta})$ up to degree-2 monomials in noise derivatives (that is, up to $O(d^2)$) and then calculate $\mathbb{E}_\pi[\cdot]$ of the result.

D.1 Expanding the Correction up to Quadratic Terms in Noise

Proposition D.1 (Expansion of the correction up to quadratic terms in noise)**.** *The additive components $L_{n,j}(\boldsymbol{\theta})/R_{n,j}(\boldsymbol{\theta})$ and $M_{n,j}(\boldsymbol{\theta})P_{n,j}(\boldsymbol{\theta})/R_{n,j}(\boldsymbol{\theta})^3$ of the correction defined in Eq. (11) admit the following formal expansion up to $O(d^2)$ and vanishing quantities as $\epsilon \to 0$:*

$$L_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1} = \big[L_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_0 + \big[L_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_1 + \big[L_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_2 + O(d^3),$$

where[1]

$$\big[L_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_0 = \frac{\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{\sqrt{g_j^2+\epsilon}}\sum_{k=0}^{n-1}\mu_{n,k}(n-k),$$

$$\big[L_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_1 = \textit{[skipped]},$$

$$\begin{aligned}
&\big[L_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_2 \\
&= \frac{1}{\sqrt{g_j^2+\epsilon}}\sum_i \frac{\partial_i g_j g_i}{2(g_i^2+\epsilon)^{5/2}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\big(3g_i^2\nu_{l,p}^2 - (g_i^2+\epsilon)\nu_{l,p} - 2(g_i^2+\epsilon)\mu_{l,p}\nu_{l,p}\big)(\partial_i d_p)^2 \\
&\quad + \frac{1}{\sqrt{g_j^2+\epsilon}}\sum_i \frac{1}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\partial_{ij}d_k\partial_i d_p \\
&\quad + \frac{1}{\sqrt{g_j^2+\epsilon}}\sum_i \frac{\partial_i g_j g_i}{(g_i^2+\epsilon)^{5/2}} \\
&\qquad\qquad \times \sum_{l=0}^{n-1}\sum_{k=0}^{l}\sum_{0\le p<q\le l}\mu_{n,k}\big(3g_i^2\nu_{l,p}\nu_{l,q} - (g_i^2+\epsilon)\mu_{l,p}\nu_{l,q} - (g_i^2+\epsilon)\mu_{l,q}\nu_{l,p}\big)\partial_i d_p\partial_i d_q \\
&\quad - \frac{g_j}{(g_j^2+\epsilon)^{3/2}}\sum_{r=0}^{n}\nu_{n,r}\sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\partial_i d_p\partial_j d_r \\
&\quad - \frac{g_j}{(g_j^2+\epsilon)^{3/2}}\sum_{r=0}^{n}\nu_{n,r}\sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}}\sum_{k=0}^{n-1}\mu_{n,k}(n-k)\partial_{ij}d_k\partial_j d_r \\
&\quad + \frac{\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{2(g_j^2+\epsilon)^{5/2}}\sum_{r=0}^{n}\big(3g_j^2\nu_{n,r}^2 - (g_j^2+\epsilon)\nu_{n,r}\big)\sum_{k=0}^{n-1}\mu_{n,k}(n-k)(\partial_j d_r)^2 \\
&\quad + \frac{3g_j^2\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{5/2}}\sum_{k=0}^{n-1}\mu_{n,k}(n-k)\sum_{0\le p<q\le n}\nu_{n,p}\nu_{n,q}\,\partial_j d_p\partial_j d_q,
\end{aligned}$$

and

$$\frac{M_{n,j}(\boldsymbol{\theta})P_{n,j}(\boldsymbol{\theta})}{R_{n,j}(\boldsymbol{\theta})^3} = \left[\frac{M_{n,j}(\boldsymbol{\theta})P_{n,j}(\boldsymbol{\theta})}{R_{n,j}(\boldsymbol{\theta})^3}\right]_0 + \left[\frac{M_{n,j}(\boldsymbol{\theta})P_{n,j}(\boldsymbol{\theta})}{R_{n,j}(\boldsymbol{\theta})^3}\right]_1 + \left[\frac{M_{n,j}(\boldsymbol{\theta})P_{n,j}(\boldsymbol{\theta})}{R_{n,j}(\boldsymbol{\theta})^3}\right]_2 + O(d^3),$$

where

$$\left[\frac{M_{n,j}(\boldsymbol{\theta})P_{n,j}(\boldsymbol{\theta})}{R_{n,j}(\boldsymbol{\theta})^3}\right]_0 := \frac{g_j^2}{(g_j^2+\epsilon)^{3/2}}\partial_j\|\boldsymbol{g}\|_{1,\epsilon}\sum_{k=0}^{n-1}(n-k)\nu_{n,k},$$

$$\left[\frac{M_{n,j}(\boldsymbol{\theta})P_{n,j}(\boldsymbol{\theta})}{R_{n,j}(\boldsymbol{\theta})^3}\right]_1 := \textit{[skipped]},$$

$$\begin{aligned}
&\left[\frac{M_{n,j}(\boldsymbol{\theta})P_{n,j}(\boldsymbol{\theta})}{R_{n,j}(\boldsymbol{\theta})^3}\right]_2 \\
&:= \frac{g_j^2\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{7/2}}\sum_{r=0}^{n-1}(n-r)\nu_{n,r}\sum_{k=0}^{n}\big(4g_j^2\nu_{n,k}^2 - (g_j^2+\epsilon)\nu_{n,k} - 2(g_j^2+\epsilon)\mu_{n,k}\nu_{n,k}\big)(\partial_j d_k)^2
\end{aligned}$$

[1] We skip the monomials of degree exactly 1 in noise derivatives because they are mean-zero and will not influence the expectation $\mathbb{E}_\pi[\cdot]$.

$$
\begin{aligned}
&+ \frac{2g_j^2\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{7/2}}\sum_{r=0}^{n-1}(n-r)\nu_{n,r} \\
&\qquad\qquad \times \sum_{0\le p<q\le n}\big(4g_j^2\nu_{n,p}\nu_{n,q}-(g_j^2+\epsilon)\mu_{n,p}\nu_{n,q}-(g_j^2+\epsilon)\mu_{n,q}\nu_{n,p}\big)\partial_j d_p\partial_j d_q \\
&+ \frac{g_j}{(g_j^2+\epsilon)^{5/2}}\sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\sum_{r=0}^{n}\nu_{n,k}\left[\mu_{l,p}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right] \\
&\qquad\qquad \times\big[(g_j^2+\epsilon)\mu_{n,r}-2g_j^2\nu_{n,r}\big]\partial_i d_p\partial_j d_r \\
&+ \frac{g_j}{(g_j^2+\epsilon)^{5/2}}\sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}}\sum_{k=0}^{n-1}\sum_{r=0}^{n}(n-k)\nu_{n,k}\big[(g_j^2+\epsilon)\mu_{n,r}-2g_j^2\nu_{n,r}\big]\partial_{ij} d_k\partial_j d_r \\
&+ \frac{\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{5/2}}\sum_{k=0}^{n-1}\sum_{r=0}^{n}(n-k)\nu_{n,k}\big[(g_j^2+\epsilon)\mu_{n,r}-2g_j^2\nu_{n,r}\big]\partial_j d_k\partial_j d_r \\
&- \frac{g_j^2\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{7/2}}\sum_{p=0}^{n-1}(n-p)\nu_{n,p}\sum_{k=0}^{n}\sum_{r=0}^{n}\nu_{n,k}\big[(g_j^2+\epsilon)\mu_{n,r}-2g_j^2\nu_{n,r}\big]\partial_j d_k\partial_j d_r \\
&+ \frac{g_j^2\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{2(g_j^2+\epsilon)^{7/2}}\sum_{r=0}^{n-1}(n-r)\nu_{n,r}\sum_{k=0}^{n}\big(3g_j^2\nu_{n,k}^2-(g_j^2+\epsilon)\nu_{n,k}\big)(\partial_j d_k)^2 \\
&+ \frac{3g_j^4\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{7/2}}\sum_{r=0}^{n-1}(n-r)\nu_{n,r}\sum_{0\le p<q\le n}\nu_{n,p}\nu_{n,q}\,\partial_j d_p\partial_j d_q \\
&- \frac{g_j^3}{(g_j^2+\epsilon)^{5/2}}\sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\left[\mu_{l,p}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\sum_{r=0}^{n}\nu_{n,r}\partial_i d_p\partial_j d_r \\
&- \frac{g_j^3}{(g_j^2+\epsilon)^{5/2}}\sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\sum_{r=0}^{n}\nu_{n,r}\partial_{ij} d_k\partial_j d_r \\
&- \frac{g_j^2\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{5/2}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\sum_{r=0}^{n}\nu_{n,r}\partial_j d_k\partial_j d_r \\
&+ \frac{g_j^2}{2(g_j^2+\epsilon)^{3/2}}\sum_i \frac{\partial_i g_j g_i}{(g_i^2+\epsilon)^{5/2}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\big(3g_i^2\nu_{l,p}^2-(g_i^2+\epsilon)\nu_{l,p}-2(g_i^2+\epsilon)\mu_{l,p}\nu_{l,p}\big)(\partial_i d_p)^2 \\
&+ \frac{g_j^2}{(g_j^2+\epsilon)^{3/2}}\sum_i \frac{\partial_i g_j g_i}{(g_i^2+\epsilon)^{5/2}} \\
&\qquad \times\sum_{l=0}^{n-1}\sum_{k=0}^{l}\sum_{0\le p<q\le l}\nu_{n,k}\big(3g_i^2\nu_{l,p}\nu_{l,q}-(g_i^2+\epsilon)\mu_{l,p}\nu_{l,q}-(g_i^2+\epsilon)\mu_{l,q}\nu_{l,p}\big)\partial_i d_p\partial_i d_q \\
&+ \frac{g_j^2}{(g_j^2+\epsilon)^{3/2}}\sum_i \frac{1}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\left[\mu_{l,p}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\partial_{ij} d_k\partial_i d_p \\
&+ \frac{g_j}{(g_j^2+\epsilon)^{3/2}}\sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\left[\mu_{l,p}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\partial_j d_k\partial_i d_p \\
&+ \frac{g_j}{(g_j^2+\epsilon)^{3/2}}\sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\partial_j d_k\partial_{ij} d_k.
\end{aligned}
$$

The proof is immediate from lemmas collected below.

Proof. To get the expansion for $L_{n,j}(\boldsymbol{\theta})/R_{n,j}(\boldsymbol{\theta})$, multiply the expansions for $L_{n,j}(\boldsymbol{\theta})$ (from Lemma D.3) and $R_{n,j}(\boldsymbol{\theta})^{-1}$ (from Lemma D.2).

To get the expansion for $M_{n,j}(\boldsymbol{\theta})P_{n,j}(\boldsymbol{\theta})/R_{n,j}(\boldsymbol{\theta})^3$, multiply the expansions for $P_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}$ and $M_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-2}$ from Lemma D.4. □

Now we state and prove the lemmas.

We start with a very simple expansion separated for pedagogical reasons to illustrate the approach (all following expansions are done similarly).

Lemma D.2 (Illustration of the approach: expansions for $M_{n,j}(\boldsymbol{\theta})$ and $R_{n,j}(\boldsymbol{\theta})^{-1}$)**.** *We have*

$$M_{n,j}(\boldsymbol{\theta}) = g_j + \sum_{k=0}^{n} \mu_{n,k} \partial_j d_k, \tag{14}$$

$$\begin{aligned} R_{n,j}(\boldsymbol{\theta})^{-1} = (g_j^2+\epsilon)^{-1/2} &- \frac{g_j}{(g_j^2+\epsilon)^{3/2}} \sum_{k=0}^{n} \nu_{n,k} \partial_j d_k \\ &+ \frac{1}{2} \sum_{k=0}^{n} \left(\frac{3 g_j^2 \nu_{n,k}^2}{(g_j^2+\epsilon)^{5/2}} - \frac{\nu_{n,k}}{(g_j^2+\epsilon)^{3/2}} \right) (\partial_j d_k)^2 \\ &+ \frac{3 g_j^2}{(g_j^2+\epsilon)^{5/2}} \sum_{0 \le p < q \le n} \nu_{n,p} \nu_{n,q} \, \partial_j d_p \partial_j d_q + O(d^3). \end{aligned}$$

Proof. Equation (14) follows directly from definitions. The expansion $R_{n,j}(\boldsymbol{\theta})^{-1}$ is obtained by the following chain of equalities:

$$\begin{aligned} R_{n,j}(\boldsymbol{\theta})^{-1} &= \left(g_j^2 + \epsilon + 2 g_j \sum_{k=0}^{n} \nu_{n,k} \partial_j d_k + \sum_{k=0}^{n} \nu_{n,k} (\partial_j d_k)^2 \right)^{-1/2} \\ &= (g_j^2+\epsilon)^{-1/2} - (g_j^2+\epsilon)^{-3/2} g_j \sum_{k=0}^{n} \nu_{n,k} \partial_j d_k - \frac{1}{2} (g_j^2+\epsilon)^{-3/2} \sum_{k=0}^{n} \nu_{n,k} (\partial_j d_k)^2 \\ &\quad + \frac{3}{2} (g_j^2+\epsilon)^{-5/2} g_j^2 \left(\sum_{k=0}^{n} \nu_{n,k} \partial_j d_k \right)^2 + O(d^3), \end{aligned}$$

where we used $\sum_{k=0}^{n} \nu_{n,k} = 1$. □

Lemma D.3 (Warm-up: expansions for $L_{n,j}(\boldsymbol{\theta})$ and $P_{n,j}(\boldsymbol{\theta})$)**.** *The following formal expansions (up to quadratic terms in noise) hold:*

$$\begin{aligned} L_{n,j}(\boldsymbol{\theta}) &= \sum_i \partial_i g_j \frac{g_i}{\sqrt{g_i^2+\epsilon}} \sum_{k=0}^{n} \mu_{n,k} (n-k) \\ &\quad + \sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}} \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \mu_{n,k} \left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon} \nu_{l,p} \right] \partial_i d_p \\ &\quad + \sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}} \sum_{k=0}^{n} \mu_{n,k} (n-k) \partial_{ij} d_k \\ &\quad + \sum_i \frac{\partial_i g_j g_i}{2(g_i^2+\epsilon)^{5/2}} \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \mu_{n,k} \big(3 g_i^2 \nu_{l,p}^2 - (g_i^2+\epsilon) \nu_{l,p} - 2(g_i^2+\epsilon) \mu_{l,p} \nu_{l,p} \big) (\partial_i d_p)^2 \\ &\quad + \sum_i \frac{1}{\sqrt{g_i^2+\epsilon}} \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \mu_{n,k} \left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon} \nu_{l,p} \right] \partial_{ij} d_k \partial_i d_p \\ &\quad + \sum_i \frac{\partial_i g_j g_i}{(g_i^2+\epsilon)^{5/2}} \\ &\qquad \times \sum_{l=0}^{n-1} \sum_{k=0}^{l} \sum_{0 \le p < q \le l} \mu_{n,k} \big(3 g_i^2 \nu_{l,p} \nu_{l,q} - (g_i^2+\epsilon) \mu_{l,p} \nu_{l,q} - (g_i^2+\epsilon) \mu_{l,q} \nu_{l,p} \big) \partial_i d_p \partial_i d_q \\ &\quad + O(d^3), \\ P_{n,j}(\boldsymbol{\theta}) &= g_j \partial_j \|\boldsymbol{g}\|_{1,\epsilon} \sum_{k=0}^{n-1} (n-k) \nu_{n,k} \end{aligned}$$

$$
\begin{aligned}
&+ g_j \sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}} \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \nu_{n,k} \left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon} \nu_{l,p} \right] \partial_i d_p \\
&+ g_j \sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}} \sum_{k=0}^{n-1} (n-k) \nu_{n,k} \partial_{ij} d_k \\
&+ \partial_j \|\boldsymbol{g}\|_{1,\epsilon} \sum_{k=0}^{n-1} (n-k) \nu_{n,k} \partial_j d_k \\
&+ g_j \sum_i \frac{\partial_i g_j g_i}{2(g_i^2+\epsilon)^{5/2}} \\
&\quad \times \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \nu_{n,k} \big(3 g_i^2 \nu_{l,p}^2 - (g_i^2+\epsilon) \nu_{l,p} - 2(g_i^2+\epsilon) \mu_{l,p} \nu_{l,p} \big) (\partial_i d_p)^2 \\
&+ g_j \sum_i \frac{\partial_i g_j g_i}{(g_i^2+\epsilon)^{5/2}} \\
&\quad \times \sum_{l=0}^{n-1} \sum_{k=0}^{l} \sum_{0 \le p < q \le l} \nu_{n,k} \big(3 g_i^2 \nu_{l,p} \nu_{l,q} - (g_i^2+\epsilon) \mu_{l,p} \nu_{l,q} - (g_i^2+\epsilon) \mu_{l,q} \nu_{l,p} \big) \partial_i d_p \partial_i d_q \\
&+ g_j \sum_i \frac{1}{\sqrt{g_i^2+\epsilon}} \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \nu_{n,k} \left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon} \nu_{l,p} \right] \partial_{ij} d_k \partial_i d_p \\
&+ \sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}} \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \nu_{n,k} \left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon} \nu_{l,p} \right] \partial_j d_k \partial_i d_p \\
&+ \sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}} \sum_{k=0}^{n-1} (n-k) \nu_{n,k} \partial_j d_k \partial_{ij} d_k \\
&+ O(d^3).
\end{aligned}
$$

Proof. By the expansion of Lemma D.2, for every $l \in [0\,{:}\,n-1]$ and coordinate i we have

$$
\begin{aligned}
&M_{l,i}(\boldsymbol{\theta}) R_{l,i}(\boldsymbol{\theta})^{-1} \\
&= \frac{g_i}{\sqrt{g_i^2+\epsilon}} + \frac{1}{\sqrt{g_i^2+\epsilon}} \sum_{p=0}^{l} \left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon} \nu_{l,p} \right] \partial_i d_p \\
&+ \frac{g_i}{2(g_i^2+\epsilon)^{5/2}} \sum_{p=0}^{l} \big(3 g_i^2 \nu_{l,p}^2 - (g_i^2+\epsilon) \nu_{l,p} - 2(g_i^2+\epsilon) \mu_{l,p} \nu_{l,p} \big) (\partial_i d_p)^2 \\
&+ \frac{g_i}{(g_i^2+\epsilon)^{5/2}} \sum_{0 \le p < q \le l} \big(3 g_i^2 \nu_{l,p} \nu_{l,q} - (g_i^2+\epsilon) \mu_{l,p} \nu_{l,q} - (g_i^2+\epsilon) \mu_{l,q} \nu_{l,p} \big) \partial_i d_p \partial_i d_q \\
&+ O(d^3).
\end{aligned}
$$

Insert this into the definition of $L_{n,j}(\boldsymbol{\theta})$. Using $\partial_{ij} \mathcal{L}_k(\boldsymbol{\theta}) = \partial_i g_j + \partial_{ij} d_k$ and keeping only terms up to degree two in the noise variables yields the claimed expansion for $L_{n,j}(\boldsymbol{\theta})$ after exchanging the order of summation and using $\sum_{l=k}^{n-1} 1 = n-k$.

Similarly, insert the same expansion into the definition of $P_{n,j}(\boldsymbol{\theta})$. Using

$$
\partial_j \mathcal{L}_k(\boldsymbol{\theta}) = g_j + \partial_j d_k, \qquad \partial_{ij} \mathcal{L}_k(\boldsymbol{\theta}) = \partial_i g_j + \partial_{ij} d_k,
$$

and truncating at quadratic order in noise gives the formula for $P_{n,j}(\boldsymbol{\theta})$. The zeroth-order term in $P_{n,j}$ is

$$
g_j \sum_i \partial_i g_j \frac{g_i}{\sqrt{g_i^2+\epsilon}} \sum_{k=0}^{n-1} (n-k) \nu_{n,k} = g_j \partial_j \|\boldsymbol{g}\|_{1,\epsilon} \sum_{k=0}^{n-1} (n-k) \nu_{n,k},
$$

since

$$
\partial_j \|\boldsymbol{g}\|_{1,\epsilon} = \sum_i \partial_j \sqrt{g_i^2+\epsilon} = \sum_i \partial_i g_j \frac{g_i}{\sqrt{g_i^2+\epsilon}}.
$$

This proves the result. □

Lemma D.4 (Preparation: expansions for $P_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}$ and $M_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-2}$)**.** *We have*

$$P_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1} = \big[P_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_0 + \big[P_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_1 + \big[P_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_2 + O(d^3),$$

where

$$\big[P_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_0 := \frac{g_j}{\sqrt{g_j^2+\epsilon}}\partial_j\|\boldsymbol{g}\|_{1,\epsilon}\sum_{k=0}^{n-1}(n-k)\nu_{n,k},$$

$$\begin{aligned}
&\big[P_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_1 \\
&:= \frac{g_j}{\sqrt{g_j^2+\epsilon}}\sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\left[\mu_{l,p}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\partial_i d_p \\
&+ \frac{g_j}{\sqrt{g_j^2+\epsilon}}\sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\partial_{ij}d_k \\
&+ \frac{\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{\sqrt{g_j^2+\epsilon}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\partial_j d_k \\
&- \frac{g_j^2\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{3/2}}\sum_{r=0}^{n-1}(n-r)\nu_{n,r}\sum_{k=0}^{n}\nu_{n,k}\partial_j d_k
\end{aligned}$$

$$\begin{aligned}
&\big[P_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}\big]_2 \\
&:= \frac{g_j\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{2(g_j^2+\epsilon)^{5/2}}\sum_{r=0}^{n-1}(n-r)\nu_{n,r}\sum_{k=0}^{n}\big(3g_j^2\nu_{n,k}^2-(g_j^2+\epsilon)\nu_{n,k}\big)(\partial_j d_k)^2 \\
&+ \frac{g_j\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{5/2}}\sum_{r=0}^{n-1}(n-r)\nu_{n,r}\sum_{0\le p<q\le n}3g_j^2\nu_{n,p}\nu_{n,q}\,\partial_j d_p\partial_j d_q \\
&- \frac{g_j^2}{(g_j^2+\epsilon)^{3/2}}\sum_i\frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\left[\mu_{l,p}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\sum_{r=0}^{n}\nu_{n,r}\partial_i d_p\partial_j d_r \\
&- \frac{g_j^2}{(g_j^2+\epsilon)^{3/2}}\sum_i\frac{g_i}{\sqrt{g_i^2+\epsilon}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\sum_{r=0}^{n}\nu_{n,r}\partial_{ij}d_k\partial_j d_r \\
&- \frac{g_j\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{3/2}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\sum_{r=0}^{n}\nu_{n,r}\partial_j d_k\partial_j d_r \\
&+ \frac{g_j}{\sqrt{g_j^2+\epsilon}}\sum_i\frac{\partial_i g_j g_i}{2(g_i^2+\epsilon)^{5/2}} \\
&\quad\times\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\big(3g_i^2\nu_{l,p}^2-(g_i^2+\epsilon)\nu_{l,p}-2(g_i^2+\epsilon)\mu_{l,p}\nu_{l,p}\big)(\partial_i d_p)^2 \\
&+ \frac{g_j}{\sqrt{g_j^2+\epsilon}}\sum_i\frac{\partial_i g_j g_i}{(g_i^2+\epsilon)^{5/2}} \\
&\quad\times\sum_{l=0}^{n-1}\sum_{k=0}^{l}\sum_{0\le p<q\le l}\nu_{n,k}\big(3g_i^2\nu_{l,p}\nu_{l,q}-(g_i^2+\epsilon)\mu_{l,p}\nu_{l,q}-(g_i^2+\epsilon)\mu_{l,q}\nu_{l,p}\big)\partial_i d_p\partial_i d_q \\
&+ \frac{g_j}{\sqrt{g_j^2+\epsilon}}\sum_i\frac{1}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\left[\mu_{l,p}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\partial_{ij}d_k\partial_i d_p
\end{aligned}$$

$$+\frac{1}{\sqrt{g_j^2+\epsilon}}\sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\left[\mu_{l,p}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\partial_j d_k \partial_i d_p$$
$$+\frac{1}{\sqrt{g_j^2+\epsilon}}\sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\partial_j d_k \partial_{ij} d_k,$$

and

$$M_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-2} = \left[M_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-2}\right]_0 + \left[M_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-2}\right]_1 + \left[M_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-2}\right]_2 + O(d^3),$$

where

$$\left[M_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-2}\right]_0 := \frac{g_j}{g_j^2+\epsilon},$$
$$\left[M_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-2}\right]_1 := \frac{1}{(g_j^2+\epsilon)^2}\sum_{k=0}^{n}\left((g_j^2+\epsilon)\mu_{n,k}-2g_j^2\nu_{n,k}\right)\partial_j d_k,$$
$$\left[M_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-2}\right]_2 := \frac{g_j}{(g_j^2+\epsilon)^3}\sum_{k=0}^{n}\left(4g_j^2\nu_{n,k}^2-(g_j^2+\epsilon)\nu_{n,k}-2(g_j^2+\epsilon)\mu_{n,k}\nu_{n,k}\right)(\partial_j d_k)^2$$
$$+\frac{2g_j}{(g_j^2+\epsilon)^3}\sum_{0\le p<q\le n}\left(4g_j^2\nu_{n,p}\nu_{n,q}-(g_j^2+\epsilon)\mu_{n,p}\nu_{n,q}-(g_j^2+\epsilon)\mu_{n,q}\nu_{n,p}\right)\partial_j d_p \partial_j d_q.$$

Proof. The expansion for $P_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-1}$ follows by multiplying the expansions for $R_{n,j}(\boldsymbol{\theta})^{-1}$ (from Lemma D.2) and $P_{n,j}(\boldsymbol{\theta})$ (from Lemma D.3).

Raising the expansion for $R_{n,j}(\boldsymbol{\theta})^{-1}$ (from Lemma D.2) to the second power yields

$$R_{n,j}(\boldsymbol{\theta})^{-2} = \frac{1}{g_j^2+\epsilon} - \frac{2g_j}{(g_j^2+\epsilon)^2}\sum_{k=0}^{n}\nu_{n,k}\partial_j d_k + \frac{1}{(g_j^2+\epsilon)^3}\sum_{k=0}^{n}\left(4g_j^2\nu_{n,k}^2-(g_j^2+\epsilon)\nu_{n,k}\right)(\partial_j d_k)^2 + \frac{8g_j^2}{(g_j^2+\epsilon)^3}\sum_{0\le p<q\le n}\nu_{n,p}\nu_{n,q}\partial_j d_p \partial_j d_q + O(d^3).$$

Multiplying this by the expansion for $M_{n,j}(\boldsymbol{\theta})$ (from Lemma D.2), we obtain the expansion for $M_{n,j}(\boldsymbol{\theta})R_{n,j}(\boldsymbol{\theta})^{-2}$, concluding the proof. □

D.2 Calculating the Expectation of the Result

Next, we calculate $\mathbb{E}_\pi[\cdot]$ of the result.

Proposition D.5 (Calculating $\mathbb{E}_\pi[\cdot]$ of the expansions obtained)**.** *We have*

$$\mathbb{E}_\pi\frac{L_{n,j}(\boldsymbol{\theta})}{R_{n,j}(\boldsymbol{\theta})} = \frac{\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{\sqrt{g_j^2+\epsilon}}\sum_{k=0}^{n-1}\mu_{n,k}(n-k)$$
$$+\frac{1}{\sqrt{g_j^2+\epsilon}}\sum_i\frac{\partial_i g_j g_i}{2(g_i^2+\epsilon)^{5/2}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\left(3g_i^2\nu_{l,p}^2-(g_i^2+\epsilon)\nu_{l,p}-2(g_i^2+\epsilon)\mu_{l,p}\nu_{l,p}\right)\mathbb{E}_\pi(\partial_i d_0)^2$$
$$+\frac{1}{\sqrt{g_j^2+\epsilon}}\sum_i\frac{1}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k=0}^{l}\mu_{n,k}\left[\mu_{l,k}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,k}\right]\mathbb{E}_\pi\partial_{ij}d_0\partial_i d_0$$

$$
\begin{aligned}
&- \frac{g_j}{(g_j^2+\epsilon)^{3/2}} \sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}} \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \mu_{n,k}\nu_{n,p} \left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right] \mathbb{E}_\pi \partial_i d_0 \partial_j d_0 \\
&- \frac{g_j}{(g_j^2+\epsilon)^{3/2}} \sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}} \sum_{k=0}^{n-1} \mu_{n,k}(n-k)\nu_{n,k}\, \mathbb{E}_\pi \partial_{ij} d_0 \partial_j d_0 \\
&+ \frac{\partial_j \|\boldsymbol{g}\|_{1,\epsilon}}{2(g_j^2+\epsilon)^{5/2}} \sum_{r=0}^{n} \big(3g_j^2\nu_{n,r}^2 - (g_j^2+\epsilon)\nu_{n,r}\big) \sum_{k=0}^{n-1} \mu_{n,k}(n-k)\, \mathbb{E}_\pi (\partial_j d_0)^2 \\
&+ O(d^3) + o_n(b^{-1})
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}_\pi \frac{M_{n,j}(\boldsymbol{\theta}) P_{n,j}(\boldsymbol{\theta})}{R_{n,j}(\boldsymbol{\theta})^3} \\
&\quad = \frac{g_j^2}{(g_j^2+\epsilon)^{3/2}} \partial_j \|\boldsymbol{g}\|_{1,\epsilon} \sum_{k=0}^{n-1} (n-k)\nu_{n,k} \\
&\qquad + \frac{g_j^2 \partial_j \|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{7/2}} \sum_{r=0}^{n-1} (n-r)\nu_{n,r} \sum_{k=0}^{n} \big(4g_j^2\nu_{n,k}^2 - (g_j^2+\epsilon)\nu_{n,k} - 2(g_j^2+\epsilon)\mu_{n,k}\nu_{n,k}\big) \mathbb{E}_\pi (\partial_j d_0)^2 \\
&\qquad + \frac{g_j}{(g_j^2+\epsilon)^{5/2}} \sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}} \\
&\qquad\quad \times \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \nu_{n,k} \left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right] \big[(g_j^2+\epsilon)\mu_{n,p} - 2g_j^2\nu_{n,p}\big] \mathbb{E}_\pi \partial_i d_0 \partial_j d_0 \\
&\qquad + \frac{g_j}{(g_j^2+\epsilon)^{5/2}} \sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}} \sum_{k=0}^{n-1} (n-k)\nu_{n,k} \big[(g_j^2+\epsilon)\mu_{n,k} - 2g_j^2\nu_{n,k}\big] \mathbb{E}_\pi \partial_{ij} d_0 \partial_j d_0 \\
&\qquad + \frac{\partial_j \|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{5/2}} \sum_{k=0}^{n-1} (n-k)\nu_{n,k} \big[(g_j^2+\epsilon)\mu_{n,k} - 2g_j^2\nu_{n,k}\big] \mathbb{E}_\pi (\partial_j d_0)^2 \\
&\qquad - \frac{g_j^2 \partial_j \|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{7/2}} \sum_{p=0}^{n-1} (n-p)\nu_{n,p} \sum_{k=0}^{n} \nu_{n,k} \big[(g_j^2+\epsilon)\mu_{n,k} - 2g_j^2\nu_{n,k}\big] \mathbb{E}_\pi (\partial_j d_0)^2 \\
&\qquad + \frac{g_j^2 \partial_j \|\boldsymbol{g}\|_{1,\epsilon}}{2(g_j^2+\epsilon)^{7/2}} \sum_{r=0}^{n-1} (n-r)\nu_{n,r} \sum_{k=0}^{n} \big(3g_j^2\nu_{n,k}^2 - (g_j^2+\epsilon)\nu_{n,k}\big) \mathbb{E}_\pi (\partial_j d_0)^2 \\
&\qquad - \frac{g_j^3}{(g_j^2+\epsilon)^{5/2}} \sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}} \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \nu_{n,k} \left[\mu_{l,p} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right] \nu_{n,p}\, \mathbb{E}_\pi \partial_i d_0 \partial_j d_0 \\
&\qquad - \frac{g_j^3}{(g_j^2+\epsilon)^{5/2}} \sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}} \sum_{k=0}^{n-1} (n-k)\nu_{n,k}^2\, \mathbb{E}_\pi \partial_{ij} d_0 \partial_j d_0 \\
&\qquad - \frac{g_j^2 \partial_j \|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{5/2}} \sum_{k=0}^{n-1} (n-k)\nu_{n,k}^2\, \mathbb{E}_\pi (\partial_j d_0)^2 \\
&\qquad + \frac{g_j^2}{2(g_j^2+\epsilon)^{3/2}} \sum_i \frac{\partial_i g_j g_i}{(g_i^2+\epsilon)^{5/2}} \\
&\qquad\quad \times \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \nu_{n,k} \big(3g_i^2\nu_{l,p}^2 - (g_i^2+\epsilon)\nu_{l,p} - 2(g_i^2+\epsilon)\mu_{l,p}\nu_{l,p}\big) \mathbb{E}_\pi (\partial_i d_0)^2 \\
&\qquad + \frac{g_j^2}{(g_j^2+\epsilon)^{3/2}} \sum_i \frac{1}{\sqrt{g_i^2+\epsilon}} \sum_{l=0}^{n-1} \sum_{k=0}^{l} \nu_{n,k} \left[\mu_{l,k} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,k}\right] \mathbb{E}_\pi \partial_{ij} d_0 \partial_i d_0 \\
&\qquad + \frac{g_j}{(g_j^2+\epsilon)^{3/2}} \sum_i \frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}} \sum_{l=0}^{n-1} \sum_{k=0}^{l} \nu_{n,k} \left[\mu_{l,k} - \frac{g_i^2}{g_i^2+\epsilon}\nu_{l,k}\right] \mathbb{E}_\pi \partial_j d_0 \partial_i d_0 \\
&\qquad + \frac{g_j}{(g_j^2+\epsilon)^{3/2}} \sum_i \frac{g_i}{\sqrt{g_i^2+\epsilon}} \sum_{k=0}^{n-1} (n-k)\nu_{n,k} \mathbb{E}_\pi \partial_{ij} d_0 \partial_j d_0 \\
&\qquad + O(d^3) + o_n(b^{-1}).
\end{aligned}
$$

Proof. Consider the average of the term like $\partial_i d_p \partial_j d_q$ where the mini-batch indices p and q are not equal:

$$\begin{aligned}
\mathbb{E}_\pi \partial_i d_p \partial_j d_q &= \mathbb{E}_\pi \frac{1}{b} \sum_{r=pb+1}^{(p+1)b} \partial_i(\ell_{\pi(r)} - \mathcal{L}) \frac{1}{b} \sum_{s=qb+1}^{(q+1)b} \partial_j(\ell_{\pi(s)} - \mathcal{L}) \\
&= \frac{1}{b^2} \sum_{r=pb+1}^{(p+1)b} \sum_{s=qb+1}^{(q+1)b} \mathbb{E}_\pi \partial_i(\ell_{\pi(r)} - \mathcal{L}) \partial_j(\ell_{\pi(s)} - \mathcal{L}) \\
&= \mathbb{E}_\pi \partial_i(\ell_{\pi(1)} - \mathcal{L}) \partial_j(\ell_{\pi(2)} - \mathcal{L}) \\
&= \frac{1}{(n+1)b((n+1)b-1)} \sum_{1 \le r_1 \ne r_2 \le (n+1)b} \partial_i(\ell_{r_1} - \mathcal{L}) \partial_j(\ell_{r_2} - \mathcal{L}) \\
&= \frac{1}{(n+1)b((n+1)b-1)} \\
&\quad \times \left(\sum_{r_1=1}^{(n+1)b} \partial_i(\ell_{r_1} - \mathcal{L}) \sum_{r_2=1}^{(n+1)b} \partial_j(\ell_{r_2} - \mathcal{L}) - \sum_{r=1}^{(n+1)b} \partial_i(\ell_r - \mathcal{L}) \partial_j(\ell_r - \mathcal{L}) \right) \\
&= -\frac{1}{(n+1)b-1} \underbrace{\frac{1}{(n+1)b} \sum_{r=1}^{(n+1)b} \partial_i(\ell_r - \mathcal{L}) \partial_j(\ell_r - \mathcal{L})}_{O(1)} \\
&= O(((n+1)b)^{-1}) = o_n(b^{-1}),
\end{aligned}$$

so, when taking expectations, we can neglect all second-degree monomials of noise derivatives where the two derivatives correspond to different mini-batches (with indices $p \ne q$ in this example). Having made this observation and recalling the expansions obtained in Proposition D.1, it is left to use the linearity of expectation and calculate basic exponential series limits:

$$\begin{aligned}
&\mathbb{E}_\pi \frac{L_{n,j}(\boldsymbol{\theta})}{R_{n,j}(\boldsymbol{\theta})} \\
&\quad = \frac{\partial_j \|\boldsymbol{g}\|_{1,\epsilon}}{\sqrt{g_j^2 + \epsilon}} \sum_{k=0}^{n-1} \mu_{n,k}(n-k) \\
&\qquad + \frac{1}{\sqrt{g_j^2 + \epsilon}} \sum_i \frac{\partial_i g_j g_i}{2(g_i^2 + \epsilon)^{5/2}} \\
&\qquad\quad \times \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \mu_{n,k} \big(3 g_i^2 \nu_{l,p}^2 - (g_i^2 + \epsilon) \nu_{l,p} - 2(g_i^2 + \epsilon) \mu_{l,p} \nu_{l,p} \big) \, \mathbb{E}_\pi (\partial_i d_0)^2 \\
&\qquad + \frac{1}{\sqrt{g_j^2 + \epsilon}} \sum_i \frac{1}{\sqrt{g_i^2 + \epsilon}} \sum_{l=0}^{n-1} \sum_{k=0}^{l} \mu_{n,k} \left[\mu_{l,k} - \frac{g_i^2}{g_i^2 + \epsilon} \nu_{l,k} \right] \mathbb{E}_\pi \partial_{ij} d_0 \partial_i d_0 \\
&\qquad - \frac{g_j}{(g_j^2 + \epsilon)^{3/2}} \sum_i \frac{\partial_i g_j}{\sqrt{g_i^2 + \epsilon}} \sum_{l=0}^{n-1} \sum_{k,p=0}^{l} \mu_{n,k} \nu_{n,p} \left[\mu_{l,p} - \frac{g_i^2}{g_i^2 + \epsilon} \nu_{l,p} \right] \mathbb{E}_\pi \partial_i d_0 \partial_j d_0 \\
&\qquad - \frac{g_j}{(g_j^2 + \epsilon)^{3/2}} \sum_i \frac{g_i}{\sqrt{g_i^2 + \epsilon}} \sum_{k=0}^{n-1} \mu_{n,k}(n-k) \nu_{n,k} \, \mathbb{E}_\pi \partial_{ij} d_0 \partial_j d_0 \\
&\qquad + \frac{\partial_j \|\boldsymbol{g}\|_{1,\epsilon}}{2(g_j^2 + \epsilon)^{5/2}} \sum_{r=0}^{n} \big(3 g_j^2 \nu_{n,r}^2 - (g_j^2 + \epsilon) \nu_{n,r} \big) \sum_{k=0}^{n-1} \mu_{n,k}(n-k) \, \mathbb{E}_\pi (\partial_j d_0)^2 \\
&\qquad + O(d^3) + o_n(b^{-1}),
\end{aligned}$$

and similarly,

$$\mathbb{E}_\pi \frac{M_{n,j}(\boldsymbol{\theta}) P_{n,j}(\boldsymbol{\theta})}{R_{n,j}(\boldsymbol{\theta})^3}$$

$$
\begin{aligned}
&= \frac{g_j^2}{(g_j^2+\epsilon)^{3/2}}\partial_j\|\boldsymbol{g}\|_{1,\epsilon}\sum_{k=0}^{n-1}(n-k)\nu_{n,k} \\
&\quad + \frac{g_j^2\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{7/2}}\sum_{r=0}^{n-1}(n-r)\nu_{n,r}\sum_{k=0}^{n}\big(4g_j^2\nu_{n,k}^2-(g_j^2+\epsilon)\nu_{n,k}-2(g_j^2+\epsilon)\mu_{n,k}\nu_{n,k}\big)\mathbb{E}_\pi(\partial_j d_0)^2 \\
&\quad + \frac{g_j}{(g_j^2+\epsilon)^{5/2}}\sum_i\frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}} \\
&\qquad \times\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\left[\mu_{l,p}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\big[(g_j^2+\epsilon)\mu_{n,p}-2g_j^2\nu_{n,p}\big]\mathbb{E}_\pi\partial_i d_0\partial_j d_0 \\
&\quad + \frac{g_j}{(g_j^2+\epsilon)^{5/2}}\sum_i\frac{g_i}{\sqrt{g_i^2+\epsilon}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\big[(g_j^2+\epsilon)\mu_{n,k}-2g_j^2\nu_{n,k}\big]\mathbb{E}_\pi\partial_{ij}d_0\partial_j d_0 \\
&\quad + \frac{\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{5/2}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\big[(g_j^2+\epsilon)\mu_{n,k}-2g_j^2\nu_{n,k}\big]\mathbb{E}_\pi(\partial_j d_0)^2 \\
&\quad - \frac{g_j^2\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{7/2}}\sum_{p=0}^{n-1}(n-p)\nu_{n,p}\sum_{k=0}^{n}\nu_{n,k}\big[(g_j^2+\epsilon)\mu_{n,k}-2g_j^2\nu_{n,k}\big]\mathbb{E}_\pi(\partial_j d_0)^2 \\
&\quad + \frac{g_j^2\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{2(g_j^2+\epsilon)^{7/2}}\sum_{r=0}^{n-1}(n-r)\nu_{n,r}\sum_{k=0}^{n}\big(3g_j^2\nu_{n,k}^2-(g_j^2+\epsilon)\nu_{n,k}\big)\mathbb{E}_\pi(\partial_j d_0)^2 \\
&\quad - \frac{g_j^3}{(g_j^2+\epsilon)^{5/2}}\sum_i\frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\left[\mu_{l,p}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,p}\right]\nu_{n,p}\,\mathbb{E}_\pi\partial_i d_0\partial_j d_0 \\
&\quad - \frac{g_j^3}{(g_j^2+\epsilon)^{5/2}}\sum_i\frac{g_i}{\sqrt{g_i^2+\epsilon}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}^2\,\mathbb{E}_\pi\partial_{ij}d_0\partial_j d_0 \\
&\quad - \frac{g_j^2\partial_j\|\boldsymbol{g}\|_{1,\epsilon}}{(g_j^2+\epsilon)^{5/2}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}^2\,\mathbb{E}_\pi(\partial_j d_0)^2 \\
&\quad + \frac{g_j^2}{2(g_j^2+\epsilon)^{3/2}}\sum_i\frac{\partial_i g_j g_i}{(g_i^2+\epsilon)^{5/2}} \\
&\qquad \times\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\big(3g_i^2\nu_{l,p}^2-(g_i^2+\epsilon)\nu_{l,p}-2(g_i^2+\epsilon)\mu_{l,p}\nu_{l,p}\big)\mathbb{E}_\pi(\partial_i d_0)^2 \\
&\quad + \frac{g_j^2}{(g_j^2+\epsilon)^{3/2}}\sum_i\frac{1}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k=0}^{l}\nu_{n,k}\left[\mu_{l,k}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,k}\right]\mathbb{E}_\pi\partial_{ij}d_0\partial_i d_0 \\
&\quad + \frac{g_j}{(g_j^2+\epsilon)^{3/2}}\sum_i\frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}\sum_{l=0}^{n-1}\sum_{k=0}^{l}\nu_{n,k}\left[\mu_{l,k}-\frac{g_i^2}{g_i^2+\epsilon}\nu_{l,k}\right]\mathbb{E}_\pi\partial_j d_0\partial_i d_0 \\
&\quad + \frac{g_j}{(g_j^2+\epsilon)^{3/2}}\sum_i\frac{g_i}{\sqrt{g_i^2+\epsilon}}\sum_{k=0}^{n-1}(n-k)\nu_{n,k}\mathbb{E}_\pi\partial_{ij}d_0\partial_j d_0 \\
&\quad + O(d^3)+o_n(b^{-1}),
\end{aligned}
$$

concluding the proof. □

Lemma D.6. *We have for all $k\in[0:n]$, $i,j\in[1:\dim\boldsymbol{\theta}]$*

$$
\begin{aligned}
\mathbb{E}_\pi\partial_i d_k\partial_j d_k &= \frac{n}{(n+1)b-1}\Sigma_{ij}, \\
\mathbb{E}_\pi\partial_{ij} d_k\partial_j d_k &= \frac{n}{2((n+1)b-1)}\partial_i\Sigma_{jj}.
\end{aligned}
$$

Proof. For any $r \in [1:(n+1)b]$ we have

$$\mathbb{E}_\pi[\partial_{ij}(\ell_{\pi(r)} - \mathcal{L})\partial_j(\ell_{\pi(r)} - \mathcal{L})] = \frac{1}{(n+1)b}\sum_{p=1}^{(n+1)b} \partial_{ij}(\ell_p - \mathcal{L})\partial_j(\ell_p - \mathcal{L}) = \frac{1}{2}\partial_i\Sigma_{jj},$$

and for $r \neq \tilde{r}$,

$$\begin{aligned}
\mathbb{E}_\pi[\partial_{ij}(\ell_{\pi(r)} - \mathcal{L})\partial_j(\ell_{\pi(\tilde{r})} - \mathcal{L})] &= \frac{1}{(n+1)b((n+1)b-1)}\sum_{\substack{p,q=1\\p\neq q}}^{(n+1)b} \partial_{ij}(\ell_p - \mathcal{L})\partial_j(\ell_q - \mathcal{L}) \\
&= -\frac{1}{(n+1)b((n+1)b-1)}\sum_{p=1}^{(n+1)b} \partial_{ij}(\ell_p - \mathcal{L})\partial_j(\ell_p - \mathcal{L}) \\
&= -\frac{1}{2((n+1)b-1)}\partial_i\Sigma_{jj}.
\end{aligned}$$

Next,

$$\begin{aligned}
\mathbb{E}_\pi\partial_i d_k\partial_j d_k &= \mathbb{E}_\pi\left(\frac{1}{b}\sum_{r=kb+1}^{kb+b}(\partial_i\ell_{\pi(r)} - \partial_i\mathcal{L})\right)\left(\frac{1}{b}\sum_{r=kb+1}^{kb+b}(\partial_j\ell_{\pi(r)} - \partial_j\mathcal{L})\right) \\
&= \frac{1}{b^2}\sum_{r=kb+1}^{kb+b}\mathbb{E}_\pi(\partial_i\ell_{\pi(r)} - \partial_i\mathcal{L})(\partial_j\ell_{\pi(r)} - \partial_j\mathcal{L}) \\
&\quad + \frac{1}{b^2}\sum_{kb+1\le r\neq\tilde{r}\le kb+b}\mathbb{E}_\pi(\partial_i\ell_{\pi(r)} - \partial_i\mathcal{L})(\partial_j\ell_{\pi(\tilde{r})} - \partial_j\mathcal{L}) \\
&= \frac{1}{b}\mathbb{E}_\pi(\partial_i\ell_{\pi(1)} - \partial_i\mathcal{L})(\partial_j\ell_{\pi(1)} - \partial_j\mathcal{L}) \\
&\quad + \frac{b-1}{b}\mathbb{E}_\pi(\partial_i\ell_{\pi(1)} - \partial_i\mathcal{L})(\partial_j\ell_{\pi(2)} - \partial_j\mathcal{L}) \\
&= \frac{1}{(n+1)b^2}\sum_{p=1}^{(n+1)b}(\partial_i\ell_p - \partial_i\mathcal{L})(\partial_j\ell_p - \partial_j\mathcal{L}) \\
&\quad + \frac{b-1}{(n+1)b^2((n+1)b-1)}\sum_{\substack{p,q=1\\p\neq q}}^{(n+1)b}(\partial_i\ell_p - \partial_i\mathcal{L})(\partial_j\ell_q - \partial_j\mathcal{L}) \\
&= \frac{1}{(n+1)b^2}\sum_{p=1}^{(n+1)b}(\partial_i\ell_p - \partial_i\mathcal{L})(\partial_j\ell_p - \partial_j\mathcal{L}) \\
&\quad - \frac{b-1}{(n+1)b^2((n+1)b-1)}\sum_{p=1}^{(n+1)b}(\partial_i\ell_p - \partial_i\mathcal{L})(\partial_j\ell_p - \partial_j\mathcal{L}) \\
&= \frac{n}{(n+1)b-1}\Sigma_{ij}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}_\pi\partial_{ij}d_k\partial_j d_k &= \frac{n}{(n+1)b-1}\frac{1}{(n+1)b}\sum_{p=1}^{(n+1)b}(\partial_{ij}\ell_p - \partial_{ij}\mathcal{L})(\partial_j\ell_p - \partial_j\mathcal{L}) \\
&= \frac{n}{2((n+1)b-1)}\partial_i\Sigma_{jj}.
\end{aligned}$$

□

Proof of Theorem B.2. Combine Proposition D.5 and Lemma D.6. □

E Proof of Theorem B.3

Proof of Assertion (a). The limit of $\mathrm{FB}_j^{(n,\epsilon)}$ follows immediately from

$$\sum_{k=0}^{n-1}\mu_{n,k}(n-k)\longrightarrow(1-\beta_1)\sum_{a=1}^{\infty}a\beta_1^a=\frac{\beta_1}{1-\beta_1},$$
$$\sum_{k=0}^{n-1}\nu_{n,k}(n-k)\longrightarrow\frac{\beta_2}{1-\beta_2}.$$

We also need the elementary limits

$$\sum_{k=0}^{n}\nu_{n,k}^2\longrightarrow\frac{1-\beta_2}{1+\beta_2},$$
$$\sum_{k=0}^{n}\mu_{n,k}\nu_{n,k}\longrightarrow\frac{(1-\beta_1)(1-\beta_2)}{1-\beta_1\beta_2},$$
$$\sum_{k=0}^{n-1}(n-k)\mu_{n,k}\nu_{n,k}\longrightarrow\frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2},$$
$$\sum_{k=0}^{n-1}(n-k)\nu_{n,k}^2\longrightarrow\frac{\beta_2^2}{(1+\beta_2)^2}.$$

Applying these four limits directly to the definition of $A_j^{(n,\epsilon)}$ gives $A_j^{(n,\epsilon)}\to A_j^{(\infty,\epsilon)}$. Since

$$\frac{n}{(n+1)b-1}=\frac{n}{(n+1)b-1}=b^{-1}+o_n(b^{-1}),$$

we obtain

$$\mathrm{MBN}_{1,j}^{(n,\epsilon)}=\mathrm{MBN}_{1,j}^{(\infty,\epsilon)}+o_n(b^{-1}).$$

Next, using the limits

$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\nu_{l,p}^2\longrightarrow\frac{\beta_1}{1-\beta_1}\frac{1-\beta_2}{1+\beta_2},$$
$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\nu_{l,p}\longrightarrow\frac{\beta_1}{1-\beta_1},$$
$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\mu_{l,p}\nu_{l,p}\longrightarrow\frac{\beta_1}{1-\beta_1}\frac{(1-\beta_1)(1-\beta_2)}{1-\beta_1\beta_2},$$
$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\nu_{l,p}^2\longrightarrow\frac{\beta_2}{1-\beta_2}\frac{1-\beta_2}{1+\beta_2},$$
$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\nu_{l,p}\longrightarrow\frac{\beta_2}{1-\beta_2},$$
$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\mu_{l,p}\nu_{l,p}\longrightarrow\frac{\beta_2}{1-\beta_2}\frac{(1-\beta_1)(1-\beta_2)}{1-\beta_1\beta_2},$$

we see that

$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\big(3g_i^2\nu_{l,p}^2-(g_i^2+\epsilon)\nu_{l,p}-2(g_i^2+\epsilon)\mu_{l,p}\nu_{l,p}\big)$$

$$\longrightarrow \frac{\beta_1}{1-\beta_1}\left(3g_i^2\frac{1-\beta_2}{1+\beta_2} - (g_i^2+\epsilon) - 2(g_i^2+\epsilon)\frac{(1-\beta_1)(1-\beta_2)}{1-\beta_1\beta_2}\right),$$

and

$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\big(3g_i^2\nu_{l,p}^2 - (g_i^2+\epsilon)\nu_{l,p} - 2(g_i^2+\epsilon)\mu_{l,p}\nu_{l,p}\big)$$
$$\longrightarrow \frac{\beta_2}{1-\beta_2}\left(3g_i^2\frac{1-\beta_2}{1+\beta_2} - (g_i^2+\epsilon) - 2(g_i^2+\epsilon)\frac{(1-\beta_1)(1-\beta_2)}{1-\beta_1\beta_2}\right).$$

Substituting these limits into $B_{i,j}^{(n,\epsilon)}$ gives

$$B_{i,j}^{(n,\epsilon)} \longrightarrow B_{i,j}^{(\infty,\epsilon)}.$$

Thus

$$\mathrm{MBN}_{2,j}^{(n,\epsilon)} = \mathrm{MBN}_{2,j}^{(\infty,\epsilon)} + o_n(b^{-1}).$$

For $\mathrm{MBN}_{3,j}^{(n,\epsilon)}$, the required scalar limits are

$$\sum_{k=0}^{n-1}(n-k)\mu_{n,k}\nu_{n,k} \longrightarrow \frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2},$$

$$\sum_{k=0}^{n-1}(n-k)\nu_{n,k} \longrightarrow \frac{\beta_2}{1-\beta_2}, \qquad \sum_{k=0}^{n-1}(n-k)\nu_{n,k}^2 \longrightarrow \frac{\beta_2^2}{(1+\beta_2)^2}.$$

Together with

$$\frac{n}{2((n+1)b-1)} = \frac{1}{2b} + o_n(b^{-1}),$$

these give

$$\mathrm{MBN}_{3,j}^{(n,\epsilon)} = \mathrm{MBN}_{3,j}^{(\infty,\epsilon)} + o_n(b^{-1}).$$

For $D_{i,j}^{(n,\epsilon)}$, we use the four limits

$$\sum_{l=0}^{n-1}\sum_{k=0}^{l}\mu_{n,k}\mu_{l,k} \longrightarrow \frac{\beta_1}{1+\beta_1},$$
$$\sum_{l=0}^{n-1}\sum_{k=0}^{l}\mu_{n,k}\nu_{l,k} \longrightarrow \frac{\beta_1(1-\beta_2)}{1-\beta_1\beta_2},$$
$$\sum_{l=0}^{n-1}\sum_{k=0}^{l}\nu_{n,k}\mu_{l,k} \longrightarrow \frac{\beta_2(1-\beta_1)}{1-\beta_1\beta_2},$$
$$\sum_{l=0}^{n-1}\sum_{k=0}^{l}\nu_{n,k}\nu_{l,k} \longrightarrow \frac{\beta_2}{1+\beta_2}.$$

Substituting them into $D_{i,j}^{(n,\epsilon)}$ gives $D_{i,j}^{(n,\epsilon)} \to D_{i,j}^{(\infty,\epsilon)}$, and the prefactor $(n)/(2((n+1)b-1)) = 1/(2b) + o_n(b^{-1})$ gives

$$\mathrm{MBN}_{4,j}^{(n,\epsilon)} = \mathrm{MBN}_{4,j}^{(\infty,\epsilon)} + o_n(b^{-1}).$$

It remains to treat $E_{i,j}^{(n,\epsilon)}$. We use the following scalar limits:

$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\nu_{n,p}\mu_{l,p} \longrightarrow \frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2},$$
$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\mu_{n,k}\nu_{n,p}\nu_{l,p} \longrightarrow \frac{\beta_1\beta_2(1-\beta_2)}{(1+\beta_2)(1-\beta_1\beta_2)},$$

$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\mu_{l,p}\mu_{n,p}\longrightarrow\frac{\beta_1\beta_2(1-\beta_1)}{(1+\beta_1)(1-\beta_1\beta_2)},$$
$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\mu_{l,p}\nu_{n,p}\longrightarrow\frac{\beta_2^2(1-\beta_1)}{(1+\beta_2)(1-\beta_1\beta_2)},$$
$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\nu_{l,p}\mu_{n,p}\longrightarrow\frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2},$$
$$\sum_{l=0}^{n-1}\sum_{k,p=0}^{l}\nu_{n,k}\nu_{l,p}\nu_{n,p}\longrightarrow\frac{\beta_2^2}{(1+\beta_2)^2},$$
$$\sum_{l=0}^{n-1}\sum_{k=0}^{l}\nu_{n,k}\mu_{l,k}\longrightarrow\frac{\beta_2(1-\beta_1)}{1-\beta_1\beta_2},$$
$$\sum_{l=0}^{n-1}\sum_{k=0}^{l}\nu_{n,k}\nu_{l,k}\longrightarrow\frac{\beta_2}{1+\beta_2}.$$

Substituting these limits into $E_{i,j}^{(n,\epsilon)}$, we get

$$\begin{aligned}
E_{i,j}^{(n,\epsilon)}\longrightarrow E_{i,j}^{(\infty,\epsilon)} :=\ & -\frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2}+\frac{g_i^2}{g_i^2+\epsilon}\frac{\beta_1\beta_2(1-\beta_2)}{(1+\beta_2)(1-\beta_1\beta_2)}\\
& -\frac{\beta_1\beta_2(1-\beta_1)}{(1+\beta_1)(1-\beta_1\beta_2)}+\frac{2g_j^2}{g_j^2+\epsilon}\frac{\beta_2^2(1-\beta_1)}{(1+\beta_2)(1-\beta_1\beta_2)}\\
& +\frac{g_i^2}{g_i^2+\epsilon}\frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2}-2\frac{g_i^2}{g_i^2+\epsilon}\frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2^2}{(1+\beta_2)^2}\\
& +\frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2^2(1-\beta_1)}{(1+\beta_2)(1-\beta_1\beta_2)}-\frac{g_i^2}{g_i^2+\epsilon}\frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2^2}{(1+\beta_2)^2}\\
& -\frac{\beta_2(1-\beta_1)}{1-\beta_1\beta_2}+\frac{g_i^2}{g_i^2+\epsilon}\frac{\beta_2}{1+\beta_2}\\
=\ & -\frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2}+\frac{g_i^2}{g_i^2+\epsilon}\frac{\beta_1\beta_2(1-\beta_2)}{(1+\beta_2)(1-\beta_1\beta_2)}\\
& -\frac{\beta_1\beta_2(1-\beta_1)}{(1+\beta_1)(1-\beta_1\beta_2)}+\frac{g_i^2}{g_i^2+\epsilon}\frac{\beta_1\beta_2(1-\beta_1)(1-\beta_2)}{(1-\beta_1\beta_2)^2}\\
& +3\frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2^2(1-\beta_1)}{(1+\beta_2)(1-\beta_1\beta_2)}-3\frac{g_i^2}{g_i^2+\epsilon}\frac{g_j^2}{g_j^2+\epsilon}\frac{\beta_2^2}{(1+\beta_2)^2}\\
& -\frac{\beta_2(1-\beta_1)}{1-\beta_1\beta_2}+\frac{g_i^2}{g_i^2+\epsilon}\frac{\beta_2}{1+\beta_2}.
\end{aligned}$$

Thus

$$E_{i,j}^{(n,\epsilon)}=E_{i,j}^{(\infty,\epsilon)}+o_n(1).$$

Since

$$\frac{n}{(n+1)b-1}=\frac{1}{b}+o_n(b^{-1}),$$

we obtain

$$\begin{aligned}
\mathrm{MBN}_{5,j}^{(n,\epsilon)} &= \frac{n}{(n+1)b-1}\frac{g_j}{g_j^2+\epsilon}\sum_i\frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}E_{i,j}^{(n,\epsilon)}\Sigma_{ij}\\
&= \frac{1}{b}\frac{g_j}{g_j^2+\epsilon}\sum_i\frac{\partial_i g_j}{\sqrt{g_i^2+\epsilon}}E_{i,j}^{(\infty,\epsilon)}\Sigma_{ij}+o_n(b^{-1})\\
&= \mathrm{MBN}_{5,j}^{(\infty,\epsilon)}+o_n(b^{-1}).
\end{aligned}$$

This proves all claimed limits. □

Proof of Assertion (b). This is obtained directly by setting $\epsilon = 0$ and simplifying the resulting expressions. The resulting limit is finite if $g_i \neq 0$ in (8). □

F Simplification of Terms

F.1 The Terms $\mathrm{MBN}_{4,j}(\beta_1, \beta_2)$ and $\mathrm{MBN}_{5,j}(\beta_1, \beta_2)$ are Small

First, the following lemma implies that the terms containing $C_4(\beta_1, \beta_2)$ and $C_5(\beta_1, \beta_2)$ are small compared to other terms and can therefore be neglected.

Lemma F.1 ($C_4(\beta_1, \beta_2)$ and $C_5(\beta_1, \beta_2)$ are small)**.** *The following bounds hold, with quotients involving C_2 interpreted by continuous extension at removable singularities:*

$$\sup_{\beta_2 \in [0.9,1)} |C_4(0.9, \beta_2)/C_1(0.9, \beta_2)| < 1.5 \times 10^{-4}, \tag{15}$$

$$\sup_{\beta_2 \in [0.9,1)} |C_5(0.9, \beta_2)/C_1(0.9, \beta_2)| < 4 \times 10^{-3}, \tag{16}$$

$$\sup_{\beta_1 \in [0.9,1)} |C_4(\beta_1, 0.999)/C_2(\beta_1, 0.999)| < 3 \times 10^{-5}, \tag{17}$$

$$\sup_{\beta_1 \in [0.9,1)} |C_5(\beta_1, 0.999)/C_2(\beta_1, 0.999)| < 5 \times 10^{-4}. \tag{18}$$

Proof. We first prove (15) and (16). Put

$$\beta_2 = \frac{9}{10} + \frac{t}{10}, \qquad t \in [0, 1).$$

On this interval,

$$C_4(0.9, \beta_2) \leq 0, \qquad C_5(0.9, \beta_2) \leq 0.$$

Therefore

$$\left| \frac{C_4(0.9, \beta_2)}{C_1(0.9, \beta_2)} \right| < \frac{3}{20000}$$

follows from

$$\frac{3}{20000} C_1(0.9, \beta_2) + C_4(0.9, \beta_2) > 0,$$

and

$$\left| \frac{C_5(0.9, \beta_2)}{C_1(0.9, \beta_2)} \right| < \frac{1}{250}$$

follows from

$$\frac{1}{250} C_1(0.9, \beta_2) + C_5(0.9, \beta_2) > 0.$$

After substituting $\beta_2 = 9/10 + t/10$, direct simplification gives

$$\frac{3}{20000} C_1(0.9, \beta_2) + C_4(0.9, \beta_2) = \frac{P_1(t)}{-1140000(t-1)(t+19)^2(9t-19)^2},$$
$$\frac{1}{250} C_1(0.9, \beta_2) + C_5(0.9, \beta_2) = \frac{P_2(t)}{-42750(t-1)(t+19)^2(9t-19)^2},$$

where

$$\begin{aligned} P_1(t) = &-26664498t^5 - 421710270t^4 + 1546038360t^3 \\ &- 1177371480t^2 + 3411450t + 178896438, \\ P_2(t) = &\ 4385502t^5 - 7335270t^4 - 324386640t^3 + 978653520t^2 - 727613550t + 178896438. \end{aligned}$$

The denominators above are positive for $t \in [0, 1)$. It remains to show that P_1 and P_2 are positive on $[0, 1)$. Using Sturm's theorem, we can observe that neither P_1 nor P_2 has a zero in $(0, 1)$. Since both are positive at 0, they are positive throughout $[0, 1]$, concluding the proof of (15) and (16).

We now treat the quotients involving C_2. Direct algebra gives

$$\frac{C_4(\beta_1,\beta_2)}{C_2(\beta_1,\beta_2)} = -\frac{(\beta_1-1)(\beta_1-\beta_2)(\beta_2-1)}{2(\beta_1+1)\left(\beta_1\beta_2^2-\beta_1\beta_2+\beta_1+\beta_2^2-2\beta_2\right)}, \tag{19}$$

$$\frac{C_5(\beta_1,\beta_2)}{C_2(\beta_1,\beta_2)} = \frac{\beta_2(\beta_1-1)(\beta_2-1)(3\beta_1-2\beta_2+1)}{(\beta_1+1)(\beta_2+1)\left(\beta_1\beta_2^2-\beta_1\beta_2+\beta_1+\beta_2^2-2\beta_2\right)}. \tag{20}$$

Put

$$\beta_1 = \frac{9}{10} + \frac{t}{10}, \qquad t \in [0,1).$$

From (19),

$$\left(\frac{3}{100000}\right)^2 - \left(\frac{C_4(\beta_1,0.999)}{C_2(\beta_1,0.999)}\right)^2 = \frac{-\left(47002997t^2-153416114t+107011917\right)\left(52997003t^2-45583886t-8011917\right)}{10000000000(t+19)^2(999001t-1008981)^2}.$$

The denominator is positive. The first quadratic factor is positive on $[0,1]$: it is decreasing on $[0,1]$, and its value at $t=1$ is positive. The second quadratic factor is negative on $[0,1]$: it is convex and negative at both endpoints. Thus the numerator is positive, giving

$$\left|\frac{C_4(\beta_1,0.999)}{C_2(\beta_1,0.999)}\right| < \frac{3}{100000} = 3\times 10^{-5}.$$

Finally, using (20),

$$\left(\frac{1}{2000}\right)^2 - \left(\frac{C_5(\beta_1,0.999)}{C_2(\beta_1,0.999)}\right)^2 = \frac{-\left(3996997001t^2-7914143962t+4316147361\right)\times\left(7991002999t^2+63938063962t-72328067361\right)}{15984004000000(t+19)^2(999001t-1008981)^2}.$$

The denominator is positive. The first quadratic factor is positive on $[0,1]$, because its discriminant is negative and its leading coefficient is positive. The second quadratic factor is negative on $[0,1]$, since it is increasing there and its value at $t=1$ is negative. Hence the numerator is positive, and

$$\left|\frac{C_5(\beta_1,0.999)}{C_2(\beta_1,0.999)}\right| < \frac{1}{2000} = 5\times 10^{-4}.$$

Equations (17) and (18) are proven. □

F.2 The Term $\mathrm{MBN}_{3,j}(\beta_1,\beta_2)$ is Neutral

Recalling

$$\mathbb{E}_\pi \partial_{ij} d_k \partial_j d_k = \frac{n}{2((n+1)b-1)} \partial_i \Sigma_{jj},$$

by Lemma D.6, we claim that

$$\begin{aligned}\mathrm{MBN}_{3,j}(\beta_1,\beta_2) &= \frac{1}{b} C_3(\beta_1,\beta_2) \frac{\operatorname{sign} g_j}{|g_j|} \sum_i \operatorname{sign} g_i \, \partial_i \Sigma_{jj} \\ &= 2C_3(\beta_1,\beta_2) \frac{\operatorname{sign} g_j}{|g_j|} \sum_i \operatorname{sign} g_i \, \mathbb{E}_\pi \partial_{ij} d_0 \partial_j d_0 + o_n(b^{-1})\end{aligned}$$

provides neither regularization nor anti-regularization, i. e. is neutral.

We start by rewriting

$$\frac{\operatorname{sign} g_j}{|g_j|} \sum_i \operatorname{sign} g_i \, \mathbb{E}_\pi \partial_{ij} d_0 \partial_j d_0 = \frac{\operatorname{sign} g_j}{|g_j|} \sum_i \operatorname{sign} g_i \, \mathbb{E}_\pi (\partial_{ij}\mathcal{L}_0 - \partial_{ij}\mathcal{L}) \partial_j d_0$$

$$= \frac{\operatorname{sign} g_j}{|g_j|} \sum_i \operatorname{sign} g_i \, \mathbb{E}_\pi \partial_{ij} \mathcal{L}_0 \partial_j d_0.$$

In the gradient-dominated (as opposed to noise-dominated) regime, the sign of a mini-batch gradient component is typically the same as the sign of the full-batch gradient component: $\operatorname{sign} \partial_i \mathcal{L}_0 \approx \operatorname{sign} \partial_i \mathcal{L}$. Then

$$\begin{aligned}\frac{\operatorname{sign} g_j}{|g_j|} \sum_i \operatorname{sign} g_i \, \mathbb{E}_\pi \partial_{ij} \mathcal{L}_0 \partial_j d_0 &\approx \frac{1}{g_j} \sum_i \mathbb{E}_\pi (\partial_j |\partial_i \mathcal{L}_0|) \partial_j d_0 \\ &= \mathbb{E}_\pi \frac{\partial_j d_0}{g_j} \partial_j \sum_i |\partial_i \mathcal{L}_0| = \mathbb{E}_\pi \frac{\partial_j d_0}{g_j} \partial_j \|\nabla \mathcal{L}_0\|_1.\end{aligned}$$

The factor $\partial_j d_0 / g_j$ (noise component relative to gradient component) can be equally likely positive or negative, so there is no preferred choice whether the 1-norm of the gradient is penalized or anti-penalized. Since our interest is the sign of (anti-)penalization, we can interpret this term as neutral for our purposes.

G Monotonicity Regions

Proposition G.1 (Monotonicity of $C_{\text{total}}(0.9, \beta_2, \lambda)$)**.** *For $\beta_2 \in [0.9, 1)$, define the auxiliary polynomials*

$$\begin{aligned}P_1(\beta_2) &:= 39753\beta_2^6 - 175145\beta_2^5 + 233419\beta_2^4 + 17703\beta_2^3 - 294872\beta_2^2 + 239758\beta_2 - 60600, \\ P_2(\beta_2) &:= 357777\beta_2^8 - 2344581\beta_2^7 + 5725125\beta_2^6 - 5463783\beta_2^5 - 1622132\beta_2^4 \\ &\quad + 8722516\beta_2^3 - 8468652\beta_2^2 + 3777388\beta_2 - 683690, \\ P_3(\beta_2) &:= 728523\beta_2^5 - 2389693\beta_2^4 + 1633232\beta_2^3 + 2272548\beta_2^2 - 3534988\beta_2 + 1289690, \\ P_4(\beta_2) &:= 2185569\beta_2^6 - 9477825\beta_2^5 + 12213510\beta_2^4 + 2832960\beta_2^3 \\ &\quad - 19805820\beta_2^2 + 16545956\beta_2 - 4500960\end{aligned}$$

and rational functions

$$\begin{aligned}f(\beta_2) &:= -\frac{(\beta_2 + 1)^3 (9\beta_2 - 10)^3}{P_1(\beta_2)}, \\ \kappa(\beta_2) &:= -\frac{(\beta_2 + 1)^4 (9\beta_2 - 10)^4}{P_2(\beta_2)}.\end{aligned}$$

Let $\rho \in (0.9, 1)$ be the unique root of $P_3(\beta_2) = 0$. (Numerically, $\rho \approx 0.9506620267$.) Set

$$\lambda_{\min} := f(\rho) \approx 0.4945366333, \qquad \lambda_{\max} := f(0.9) = \frac{6859}{13497} \approx 0.5081870045,$$

and

$$\kappa_{\min} := \kappa(0.9) = \frac{130321}{271013} \approx 0.4808662315.$$

Then the monotonicity of $C_{\text{total}}(0.9, \beta_2, \lambda)$, as a function of $\beta_2 \in [0.9, 1)$, is as follows.

(a) If $\lambda \le \lambda_{\min}$, then $C_{\text{total}}(0.9, \beta_2, \lambda)$ is strictly decreasing on $[0.9, 1)$.

(b) If $\lambda_{\min} < \lambda < 1/2$, then there are unique points

$$u_\lambda \in (0.9, \rho), \qquad v_\lambda \in (\rho, 1)$$

such that

$$f(u_\lambda) = f(v_\lambda) = \lambda.$$

Moreover,

$$\begin{aligned}&C_{\text{total}}(0.9, \beta_2, \lambda) \text{ is decreasing on } [0.9, u_\lambda), \\ &C_{\text{total}}(0.9, \beta_2, \lambda) \text{ is increasing on } (u_\lambda, v_\lambda), \\ &C_{\text{total}}(0.9, \beta_2, \lambda) \text{ is decreasing on } (v_\lambda, 1).\end{aligned}$$

(c) If $1/2 \le \lambda < \lambda_{\max}$, then there is a unique point

$$u_\lambda \in (0.9, \rho)$$

such that

$$f(u_\lambda) = \lambda.$$

Moreover,

$$C_{\text{total}}(0.9, \beta_2, \lambda) \text{ is decreasing on } [0.9, u_\lambda),$$
$$C_{\text{total}}(0.9, \beta_2, \lambda) \text{ is increasing on } (u_\lambda, 1).$$

(d) If $\lambda \ge \lambda_{\max}$, then $C_{\text{total}}(0.9, \beta_2, \lambda)$ is increasing on $[0.9, 1)$.

Proof. Define

$$S(\beta_2) := C_1(0.9, \beta_2) + C_2(0.9, \beta_2).$$

Then

$$C_{\text{total}}(0.9, \beta_2, \lambda) = 9 - \frac{\beta_2}{1-\beta_2} + \lambda S(\beta_2).$$

A direct simplification and differentiation gives

$$S(\beta_2) = -\frac{3672\beta_2^5 - 8905\beta_2^4 + 2677\beta_2^3 + 8282\beta_2^2 - 7428\beta_2 + 1710}{(\beta_2 - 1)(\beta_2 + 1)^2(9\beta_2 - 10)^2},$$
$$S'(\beta_2) = -\frac{P_1(\beta_2)}{(\beta_2 - 1)^2(\beta_2 + 1)^3(9\beta_2 - 10)^3}.$$

We first record the sign information needed below. Using Sturm's theorem, we can see that P_1 has no root in $(0.9, 1)$ and is positive there, while P_2 has no root in $(0.9, 1)$ and is negative there. Moreover, P_3 has exactly one root in $(0.9, 1)$, denoted by ρ, and P_4 has no root in $(0.9, 1)$ and is negative there.

Since

$$(\beta_2 - 1)^2(\beta_2 + 1)^3(9\beta_2 - 10)^3 < 0$$

on $[0.9, 1)$, and since $P_1(\beta_2) > 0$, we have

$$S'(\beta_2) > 0 \qquad \text{for } \beta_2 \in [0.9, 1).$$

Now

$$\begin{aligned}\partial_{\beta_2} C_{\text{total}}(0.9, \beta_2, \lambda) &= -\frac{1}{(1-\beta_2)^2} + \lambda S'(\beta_2) \\ &= S'(\beta_2)\left(\lambda - \frac{1}{(1-\beta_2)^2 S'(\beta_2)}\right) \\ &= S'(\beta_2)\left(\lambda - f(\beta_2)\right).\end{aligned}$$

Because $S'(\beta_2) > 0$, the sign of

$$\partial_{\beta_2} C_{\text{total}}(0.9, \beta_2, \lambda)$$

is the sign of

$$\lambda - f(\beta_2).$$

We next analyze f. Direct differentiation gives

$$f'(\beta_2) = \frac{2(\beta_2 - 1)(\beta_2 + 1)^2(9\beta_2 - 10)^2 P_3(\beta_2)}{P_1(\beta_2)^2}.$$

Since P_3 has a unique root $\rho \in (0.9, 1)$, with

$$P_3(\beta_2) > 0 \text{ on } [0.9, \rho), \qquad P_3(\beta_2) < 0 \text{ on } (\rho, 1),$$

and since $\beta_2 - 1 < 0$, we get

$$f'(\beta_2) < 0 \text{ on } [0.9, \rho), \qquad f'(\beta_2) > 0 \text{ on } (\rho, 1).$$

Therefore f decreases on $[0.9, \rho]$ and increases on $[\rho, 1)$. Furthermore,

$$f(0.9) = \frac{6859}{13497}, \qquad \lim_{\beta_2 \to 1^-} f(\beta_2) = \frac{1}{2}.$$

Thus $f(\rho) = \lambda_{\min}$ is the minimum of f, and the largest value of f on $[0.9, 1)$ is $f(0.9) = \lambda_{\max}$.

The monotonicity classification follows from the sign rule

$$\operatorname{sign} \partial_{\beta_2} C_{\text{total}}(0.9, \beta_2, \lambda) = \operatorname{sign}(\lambda - f(\beta_2)),$$

together with the fact that f decreases from $\lambda_{\max}$ to $\lambda_{\min}$, then increases from $\lambda_{\min}$ to $1/2$. □

Proposition G.2 (Monotonicity of $C_{\text{total}}(\beta_1, 0.999, \lambda)$)**.** *For $\beta_1 \in [0.9, 1)$, define a polynomial*

$$P_1(\beta_1) := 2001994001001\beta_1^3 - 6011966022994\beta_1^2 + 6017956024997\beta_1 - 2007984005000$$

and rational function

$$f(\beta_1) := \frac{1999(999\beta_1 - 1000)^3}{P_1(\beta_1)}.$$

Set

$$\lambda_{\min}^{(1)} := f(0.9) = \frac{2053460214271}{2062434398111} \approx 0.9956487422,$$
$$\lambda_{\max}^{(1)} := \lim_{\beta_1 \to 1^-} f(\beta_1) = \frac{1999}{1996} \approx 1.001503006.$$

Then the monotonicity of $C_{\text{total}}(\beta_1, 0.999, \lambda)$ as a function of $\beta_1 \in [0.9, 1)$ is as follows.

(a) If $\lambda \le \lambda_{\min}^{(1)}$, then $C_{\text{total}}(\beta_1, 0.999, \lambda)$ is increasing on $[0.9, 1)$.

(b) If $\lambda_{\min}^{(1)} < \lambda < \lambda_{\max}^{(1)}$, then there is a unique point

$$u_\lambda \in (0.9, 1)$$

such that

$$f(u_\lambda) = \lambda.$$

Moreover,

$$C_{\text{total}}(\beta_1, 0.999, \lambda) \text{ is decreasing on } [0.9, u_\lambda),$$

and

$$C_{\text{total}}(\beta_1, 0.999, \lambda) \text{ is increasing on } (u_\lambda, 1).$$

(c) If $\lambda \ge \lambda_{\max}^{(1)}$, then $C_{\text{total}}(\beta_1, 0.999, \lambda)$ is strictly decreasing on $[0.9, 1)$.

Proof. Write

$$S(\beta_1) := C_1(\beta_1, 0.999) + C_2(\beta_1, 0.999).$$

Then

$$C_{\text{total}}(\beta_1, 0.999, \lambda) = \frac{\beta_1}{1 - \beta_1} - \frac{0.999}{1 - 0.999} + \lambda S(\beta_1) = \frac{\beta_1}{1 - \beta_1} - 999 + \lambda S(\beta_1).$$

A direct simplification and differentiation gives

$$S(\beta_1) = \frac{7973042963014998\beta_1^3 - 23931084945018997\beta_1^2 + 23943040997990003\beta_1 - 7984999011996000}{3996001(\beta_1 - 1)(999\beta_1 - 1000)^2},$$
$$S'(\beta_1) = -\frac{P_1(\beta_1)}{1999(\beta_1 - 1)^2(999\beta_1 - 1000)^3}.$$

We first determine the sign of $S'(\beta_1)$ on $[0.9, 1)$. Since the denominator

$$1999(\beta_1 - 1)^2(999\beta_1 - 1000)^3$$

is negative for $\beta_1 \in [0.9, 1)$,

$$\operatorname{sign} S'(\beta_1) = \operatorname{sign} P_1(\beta_1).$$

We claim that

$$P_1(\beta_1) < 0 \qquad \text{for all } \beta_1 \in [0.9, 1]. \tag{21}$$

To see this, put

$$\beta_1 = \frac{9}{10} + \frac{t}{10}, \qquad t \in [0, 1].$$

Then

$$\begin{aligned} &1000\, P_1\left(\frac{9}{10} + \frac{t}{10}\right) \\ &\quad = 2001994001001t^3 - 6065822202913t^2 + 6126260604023t - 2062434398111. \end{aligned}$$

Differentiating, one obtains that the right-hand side increases in $t \in [0, 1]$. Since the value at $t = 1$ is negative, (21) is true. We have proven that

$$S'(\beta_1) < 0 \qquad \text{for all } \beta_1 \in [0.9, 1).$$

Now differentiate C_{total}:

$$\partial_{\beta_1} C_{\text{total}}(\beta_1, 0.999, \lambda) = \frac{1}{(1-\beta_1)^2} + \lambda S'(\beta_1) = S'(\beta_1)\big(\lambda - f(\beta_1)\big),$$

where

$$f(\beta_1) := -\frac{1}{(1-\beta_1)^2 S'(\beta_1)} = \frac{1999(999\beta_1 - 1000)^3}{P_1(\beta_1)}.$$

Since $S'(\beta_1) < 0$, this implies

$$\operatorname{sign} \partial_{\beta_1} C_{\text{total}}(\beta_1, 0.999, \lambda) = \operatorname{sign}\big(f(\beta_1) - \lambda\big).$$

It remains to analyze f. Direct differentiation gives

$$f'(\beta_1) = \frac{7992002(\beta_1 - 1)(999\beta_1 - 1000)^2(6990003\beta_1 - 6994000)}{P_1(\beta_1)^2}.$$

On $[0.9, 1)$, we have

$$\beta_1 - 1 < 0, \qquad (999\beta_1 - 1000)^2 > 0, \qquad 6990003\beta_1 - 6994000 < 0.$$

Since $P_1(\beta_1)^2 > 0$, it follows that

$$f'(\beta_1) > 0 \qquad \text{for all } \beta_1 \in [0.9, 1).$$

Thus f is strictly increasing on $[0.9, 1)$. Moreover,

$$f(0.9) = \frac{2053460214271}{2062434398111} = \lambda^{(1)}_{\min},$$

and

$$\lim_{\beta_1 \to 1^-} f(\beta_1) = \frac{1999}{1996} = \lambda^{(1)}_{\max}.$$

We now translate this into monotonicity of $C_{\text{total}}(\beta_1, 0.999, \lambda)$. The sign rule is

$$\operatorname{sign} \partial_{\beta_1} C_{\text{total}}(\beta_1, 0.999, \lambda) = \operatorname{sign}\big(f(\beta_1) - \lambda\big).$$

If $\lambda \le \lambda^{(1)}_{\min}$, then

$$f(\beta_1) - \lambda \ge 0 \qquad \text{for all } \beta_1 \in [0.9, 1),$$

and the inequality is strict for $\beta_1 > 0.9$. Hence $C_{\text{total}}(\beta_1, 0.999, \lambda)$ is increasing on $[0.9, 1)$, and strictly increasing on $(0.9, 1)$.

If

$$\lambda_{\min}^{(1)} < \lambda < \lambda_{\max}^{(1)},$$

then, because f is continuous and strictly increasing, there is a unique

$$u_\lambda \in (0.9, 1)$$

such that

$$f(u_\lambda) = \lambda.$$

For $\beta_1 < u_\lambda$, we have $f(\beta_1) < \lambda$, so

$$\partial_{\beta_1} C_{\text{total}}(\beta_1, 0.999, \lambda) < 0.$$

For $\beta_1 > u_\lambda$, we have $f(\beta_1) > \lambda$, so

$$\partial_{\beta_1} C_{\text{total}}(\beta_1, 0.999, \lambda) > 0.$$

Therefore $C_{\text{total}}(\beta_1, 0.999, \lambda)$ is decreasing on $[0.9, u_\lambda)$ and increasing on $(u_\lambda, 1)$.

Finally, if $\lambda \geq \lambda_{\max}^{(1)}$, then

$$f(\beta_1) - \lambda < 0 \qquad \text{for all } \beta_1 \in [0.9, 1),$$

because $f(\beta_1) < \lambda_{\max}^{(1)}$ for every $\beta_1 < 1$. Hence

$$\partial_{\beta_1} C_{\text{total}}(\beta_1, 0.999, \lambda) < 0$$

throughout $[0.9, 1)$, and $C_{\text{total}}(\beta_1, 0.999, \lambda)$ is strictly decreasing on $[0.9, 1)$. This proves the classification. □